Supplemental Materials for GEViTRec

April 16, 2021

Contents

S1 Implementation Details

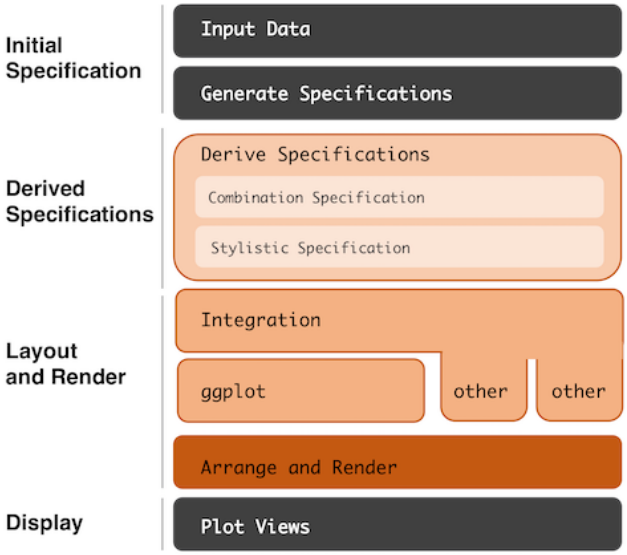


Figure S1: GEViTRec API Architecture

We describe the proof-of-concept implementation of our recommendation algorithm, which provides evidence for the viability of our approach and allows us to conduct a preliminary evaluation of its utility.

S1.1 Language and Environment

We chose to implement `GEViTRec` as a package in the R programming language. We chose that language to fit into the workflows of genEpi experts: they routinely use R to perform statistical analyses, and integration within it allows them to use that environment to further filter or refine their analysis. The static combinations of charts that are created by `GEViTRec` are also usable for communication with others in addition to supporting the data recon process, since PDF reports remain the primary medium of communication between genEpi experts [1].

We thus build on ggplot [2] as the charting library layer for many of the statistical charts, and a variety of other R packages such as ggtree [3] for more specialized charts.

Our package of code, evaluation data, and documentation is online: https://github.com/amcrisan/GEViTRec

S1.2 GEViTRec API

The Application Programming Interface (API) for the `GEViTRec` package is designed for extreme simplicity, in keeping with the goal of a recommender system suitable for data recon. It is intended to be run from an R environment as a set of functions to demonstrate the concept; it could be transformed into a standalone interactive application as future work. Figure S1 illustrates the software architecture of our implementation. The API has functions to help users load heterogeneous data, perform data integration, generate specifications for chart combinations, and render those specifications as displayable plots.

`input_data` is a common interface for loading different data types, through a series of data-type-specific functions that we developed to load and store datasets in a standardized format.

`data_linkage` takes a collection of datasets and explodes attribute fields from different data sources, finds linkages, and creates the data source graph. The `view_entity_graph` command will display that graph. Users can also view a metadata table for the different data sources and their attribute fields.

`get_spec_list` performs computations on the data source graph, ranking paths and processing them in order of rank (highest to lowest) to generate chart specifications.

`plot_view` takes a list of chart specifications, renders them, and arranges the resulting pixels together into a view: a single large box of displayable pixels, containing multiple visually coherent static charts. A user parameter indicates which of the ranked views to display.

Separating the data integration and specification computation into separate functions allows the optional display of the data source graph data structure itself, if desired by the user. If a data source 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 static combination to five, which are selected based upon their relevance scores. These choices are easily modifiable.

S2 Survey Procedures

Consent and questionnaire administered in the `GEViTRec` User assessment (begins on following page). This study was approved by the [redacted] Behavioral Ethics Research Board; certificate number: H10-03336

Informed Consent

GEViTRec User Study

DESCRIPTION AND CONSENT

People use a variety of systems to develop data visualizations. These systems vary in their ease of use and their abilities to support the visualization needs of different users. From previous research we specifically found that stakeholders can become overwhelmed by new and large collections of data and that they need help visualizing these data. We have developed GEViTRec, an R-based data visualization tool that automatically visualizes data in order to help stakeholders in genomic epidemiology quickly view and assess their data.

Given a set of input data, GEViTRec is able to automatically produce an ordered set of data visualization suggestions. GEViTRec is able to do so because it builds upon prior results from a Genomic Epidemiology Visualization Typology (GEViT) that our research group developed through a systematic analysis of data visualization practices in genomic epidemiology. Combining information about stakeholders input data with the results of the GEViT research, our GEViTRec system is able to identify and prioritize the kinds of data visualizations that are relevant to stakeholders.

In this user study, we seek to assess the usefulness of GEViTRec's data visualization outputs. By participating in this user study, your will help us better understand the strengths and weakness of GEViTRec and how data visualization systems in general can better support users.

STUDY PROCEDURES:

We estimate that this study will take approximately 20 – 30 minutes total to complete the GEViTRec presentation and online survey.

There are no known or anticipated risks to you by participating in this research. Study results will be shared with the research community through open- access publications, conference reports, tweets, and other social media postings. Anonymized transcripts and survey responses will also be publicly shared as part of the supplementary materials for subsequent publications.

MEASURES TO MAINTAIN CONFIDENTIALITY

You are assigned a unique numeric identifier and we do not collect any data that links you specifically to this participant identifier number. Data from this study will be coded anonymously and no identifiable personal information is collected.

All survey data is collected through [redacted] Qualtrics. The Qualtrics platform compiles with the [redacted] Freedom of Information and Protection of Privacy Act (FIPPA), which keeps data secure, backed-up, and stored in [redacted]

CONTACTS FOR COMPLAINTS OR CONCERNS

[redacted]

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 and to have your anonymized results shared online following the completion of the study.

PRINCIPAL INVESTIGATOR

[redacted]

CO-INVESTIGATORS

[redacted]

- ○ I consent to participate in the study
- ○ I do not consent to participate in the study

Demographic Questions

Participant Background

This set of questions asks for your background, your programming expertise, and about the data visualization tools that you use

Please enter the unique study identifier that has been provided to you

Please indicate your role in infectious disease diagnosis, treatment, management, and/or surveillance. You may select more than one option

- ☐ Clinical management - I work directly with patients, providing care and/or case management

[redacted]

- ☐ Laboratory work - I work in a lab where I am involved with testing for different pathogens
- ☐ Surveillance Analyses/ Data Science - I work with data to understand disease occurrence
- ☐ Bioinformatics - I primarily analyze genomic data
- ☐ Epidemiologist / Biostatistician - I have specialized knowledge and analyze public health data
- ☐ Other (please specify) _____

How would you rate your ability to program in R?

- ○ No ability - I have never used R and am not confident I could pick it up
- ○ Possible ability - I program in other languages and feel I could pick up R
- ○ Beginner - I program in R every now and then
- ○ Intermediate - I can confidently carry out analyses in R
- ○ Expert - I develop packages in R

How do you rate your ability of program in general?

- ○ No ability - I never write code
- ○ Beginner - I can write some code
- ○ Intermediate - I regularly write code for analyses or small applications
- ○ Expert - I can code complex applications that others use

How frequently do you produce data visualizations?

- ○ Never or Rarely
- ○ Occasionally - Usually only for paper figures
- ○ Frequently - Use it to understand and communicate my results

Please indicate how often you use these difference data visualization tools. Use your judgement, responses don't have to be very rigid (i.e. if you don't use something literally every single day, but use it pretty often for your work, select the daily basis option)

	I never use this	A few times a year	Monthly basis	Daily basis
R (ggplot, tidyverse, etc.)	○	○	○	○
Python (matplot; bokeh, seaborn, etc.)	○	○	○	○
Java Script (D3; Vega; Vega-lite; etc.)	○	○	○	○
Tableau	○	○	○	○
Excel	○	○	○	○
Google Docs	○	○	○	○

[Optional] What other data visualization systems or programming libraries that you use regularly (daily or weekly basis), but that are not listed in the previous question. Leave blank if you have no response.

Chauffered Demo

GEViTRec Demonstration

You will be shown a demonstration of the GEViTRec system using two different datasets. One dataset is synthetic, the other dataset is publicly available data from the 2014 - 2016 Ebola Outbreak. The survey administrator will ask you a set of pre-specified questions during this demonstration. Your responses will be recorded via a voice recording device and the administrator will also take notes.

Transcripts and notes of your responses will be anonymized and made publicly available. **Audio recordings of the discussion will not be released**.

- O I consent to transcripts and summary notes of the discussion being released
- O I only consent to summary notes of the discussion being released
- O I no longer wish to participate

GEViT Rec Assessment Questions

GEViTRec Assessment

This set of questions will ask you about your session with GEViTRec

Please answer the following questions about the session with GEViTRec

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
It was easy to import data	O	O	O	O	O
It was easy to understand how the data connected together	O	O	O	O	O
I could quickly find a useful data visualization from the set of suggestions provided by the system	O	O	O	O	O
The ordering of the suggestions made sense to me	O	O	O	O	O
It was easier for me to pick one of the suggestions than make a visualization on my own	O	O	O	O	O

[redacted]

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
The visualization suggestions included things I did not think of myself	O	O	O	O	O
Some of the things I thought of were not included in the suggested visualizations	O	O	O	O	O

Please answer the following questions on visualizations produced by GEViTRec

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strong disagree
The visualizations that I saw were relevant to Genomic Epidemiology	O	O	O	O	O
The visualizations helped me understand the data	O	O	O	O	O
I would share these visualizations with others	O	O	O	O	O
The way the visualization suggestions changed according to the dataset seemed appropriate	O	O	O	O	O
These visualizations would help my current analyses	O	O	O	O	O

Please briefly describe the features of GEViTRec that you found to be most useful:

Please briefly describe the features of GEViTRec that you found to be confusing or not helpful

EBOV DATA

GEViTRec Comparison to Existing Systems

This set of questions will ask you to compare GEViTREC's suggestions to other data visualizations. The visualizations you will be asked to compare are generated from publicly available datasets from the 2014 - 2016 Ebola outbreak.

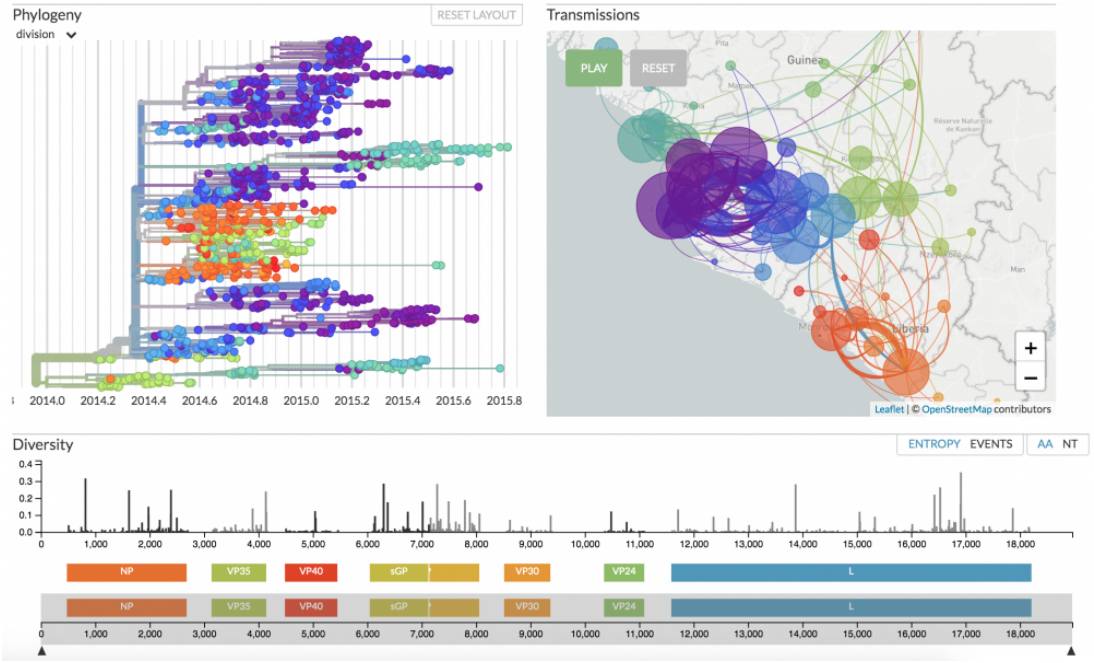
Have you used the Nextstrain or Microreact to visualize your own data or to explore some new dataset?

O I have previously used Nextstrain and/or Microreact

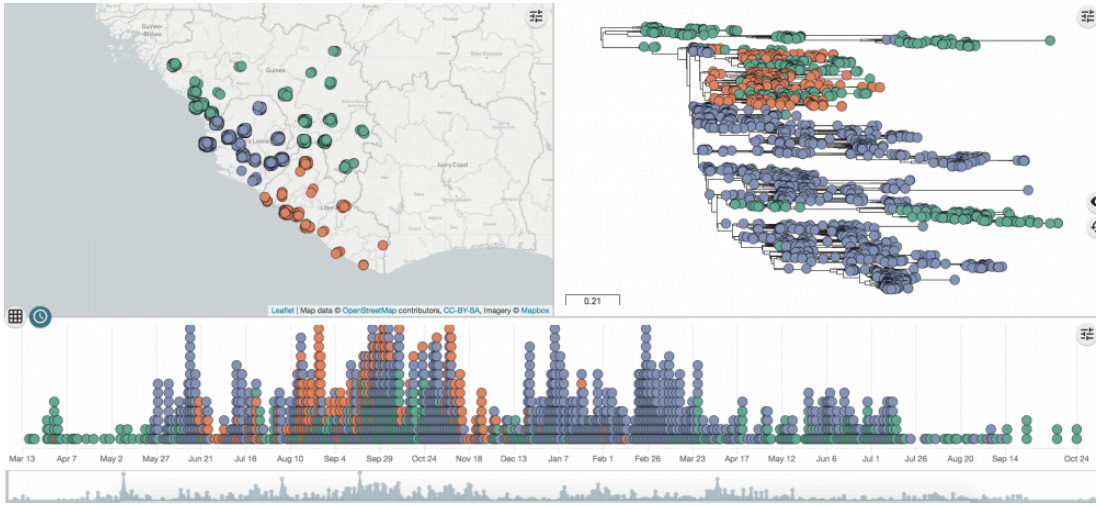
- ○ I have not previously used Nextstrain and/or Microreact, but I am aware of them
- ○ I do not know what these are

The next set of questions as your compare the the top ranked visualization automatically produced by GEViTRec against the (custom designed) visualizations of Nextstrain and Microreact. Please click on the images to see a larger view

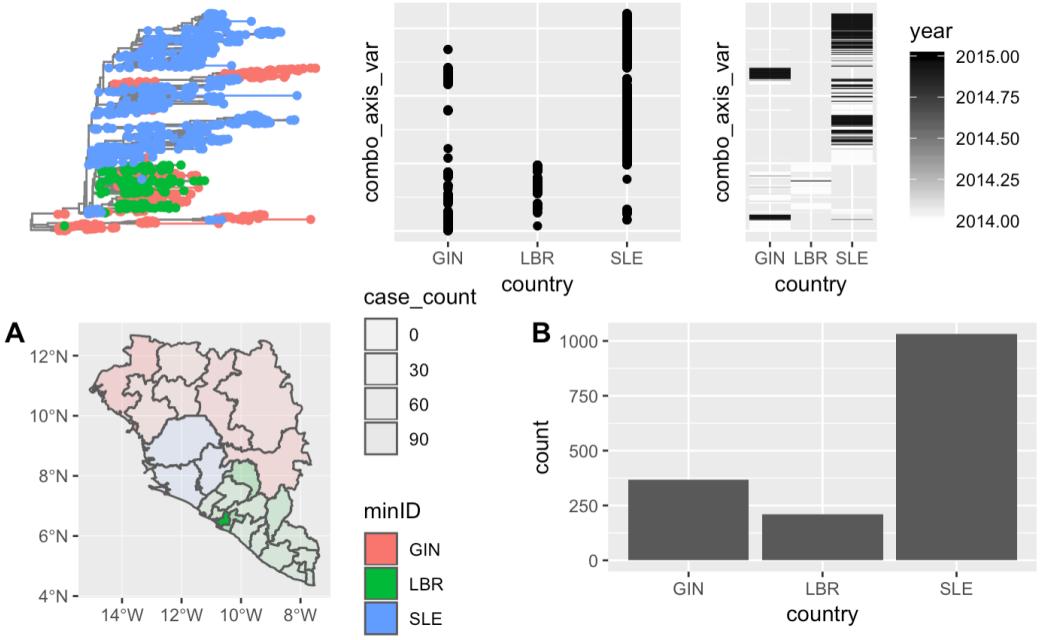
A) Nextstrain (www.nextstrain.org)



B) Microreact (https://microreact.org/)



C) GEViTRec



How long do you think it would take you to create the data visualizations shown in the Nextstrain, Microreact, and GEViTRec systems?

	I can't generate this visualization	A few hours or less	A few days	Weeks
Nextstrain	O	O	O	O
Microreact	O	O	O	O

[redacted]

	I can't generate this visualization	A few hours or less	A few days	Weeks
GEViTRec	O	O	O	O

Please state the extent to which you agree with the following features of data visualization systems:

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
Data visualizations systems should integrate directly with many different analysis tools	O	O	O	O	O
Data visualization systems need to have interactive components	O	O	O	O	O
Data visualization systems should help me make visualizations	O	O	O	O	O
Data visualization systems should let me pick the visualizations I care most about	O	O	O	O	O
Data visualization systems should show the same visual representations even if the input datasets are different	O	O	O	O	O

[OPTIONAL] Is there anything else you'd like to tell us about the comparison of Nextstrain, Microreact, and GEViTRec that we have not asked you about? Leave blank if there's nothing else.

GEViTRec Own Data

GEViTRec Usage

The next few questions will ask you about using GEViTRec on your own data now or in the future

Would use GEViTRec on your own datasets?

- O Yes
- O No
- O I already have!
- O Don't Know
- O Other (please specify):

[OPTIONAL] Can you briefly describe the type of data you used or that you might use in the future? (i.e how many datasets, how big are the datasets, what types of data do they contain).

[redacted]

Leave blank if you are sure that you would not use GEViTRec on your own data.

[OPTIONAL] If you used GEViTRec, please tell us about your experience using GEViTRec (i.e. did it help you understand your data or share it with colleagues).

Leave blank if you have not used GEViTRec on your own data.

What do you think could be improved about GEViTRec?

Is there anything else you'd like to tell us about your experience using GEViTRec that we have not asked you about? Leave blank if there's nothing else.

S3 Survey Demos: Synthetic Data and Ebola Outbreak data

Here we present the R Markdown documents that were used for the chauffeured demonstration. In these documents, all areas in gray text are R code that was run, where as all areas in white are text or results. Note that participants were not walked through a PDF document; we ran the markdown code so participants could observe how long it takes to generate the results.

Note that there is some distortion of the image here because they have been sized to fit the page.

Synthetic Data GEViTRec

```
devtools::load_all() #temporary once things are done

set.seed(416)

library(mincombinr)
library(dplyr)
library(shiny)
library(treeio)

tab_dat<-input_data(file = "../extdata/syn_tab_dat.csv",
                    dataType = "table")

tree_dat<-input_data(file = system.file("extdata", "sample.nwk",
                                        package="treeio"),
                     dataType = "tree")

genomic_dat<-input_data(file = system.file("extdata", "synth.fasta",
                                           package = "mincombinr"),
                        dataType = "dna")

gel_img <- input_data(file = system.file("extdata", "synth_gel_image.tiff",
                                         package = "mincombinr"),
                              dataType = "image")


id<-tree_dat@data$tipData
```

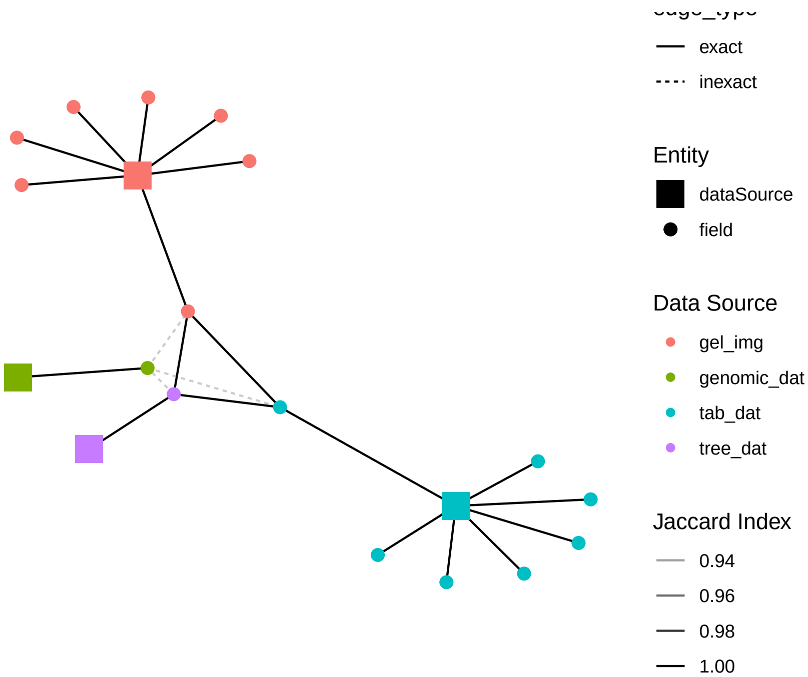
Adding annotations for the image data

```
#not evaluated
gel_img_tmp<-gel_img #make a copy of get image

gel_img_tmp <- annotate_image(gel_img) #annotating the image
```

```
#added pre-loaded data
tmp<-readRDS(file="../extdata/syn_gel_image_meta.rds")

gel_img@data$metadata<-tmp
```

Data Harmonization

```
harmon_obj<-data_harmonization(tab_dat,tree_dat,
                               genomic_dat,gel_img)

#plotting the entity graph
view_entity_graph(harmon_obj[["entityGraph"]])
```



Generate Specifications

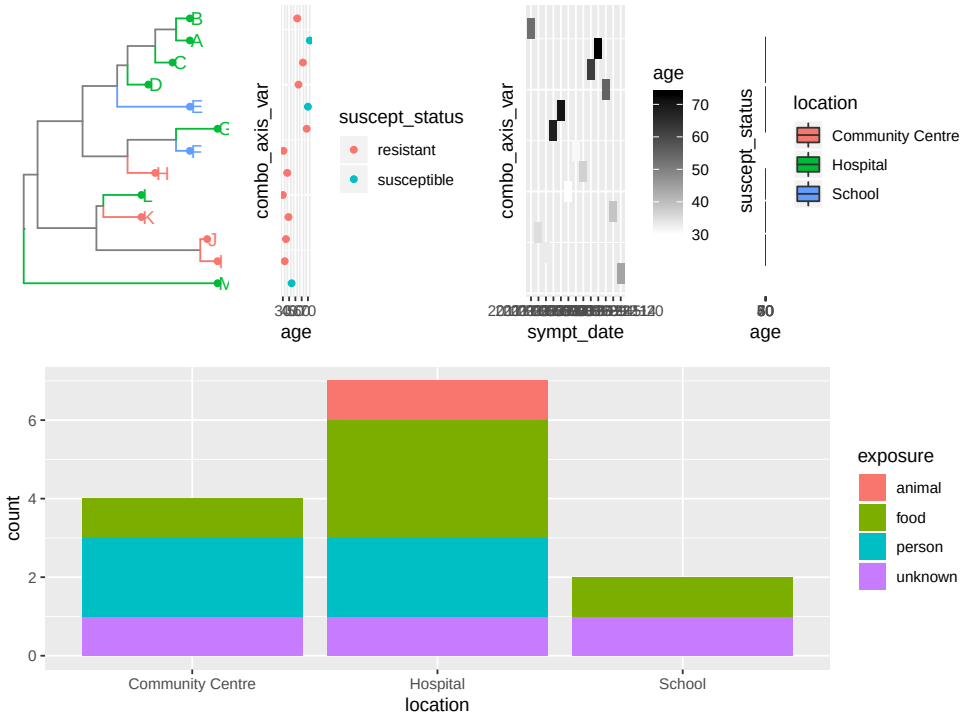
```
component_specs<-get_spec_list(harmon_obj)
```

```
## [1] "Generating possible specifications"
## [1] "Cleaning specifications up"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
```

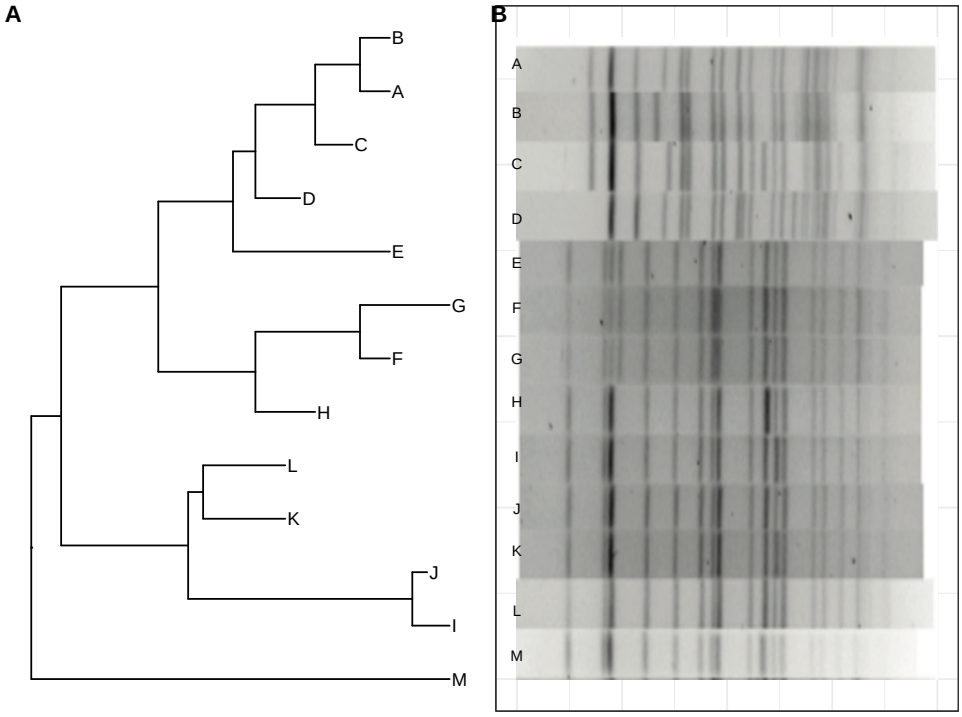
```
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
```

Look at the generated views

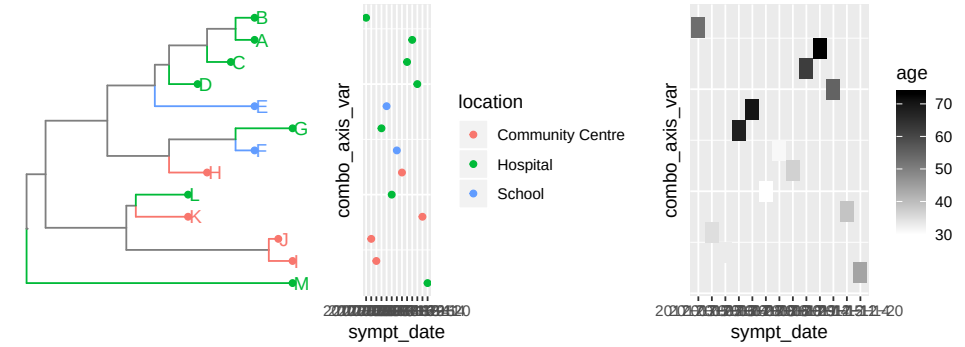
```
plot_view(component_specs,view_num=1)
```

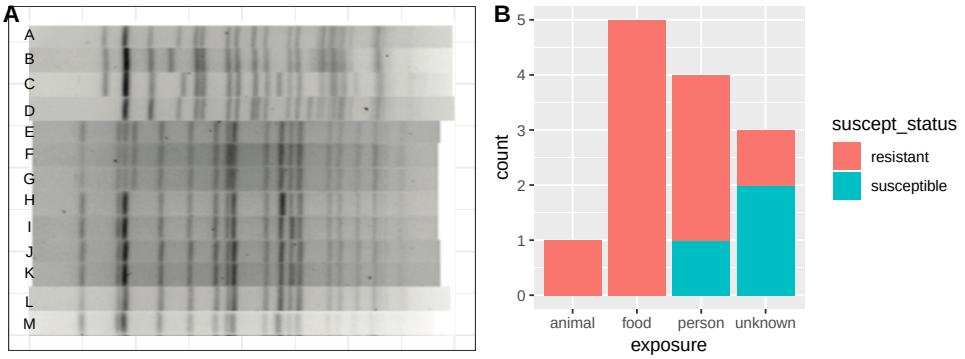



```
plot_view(component_specs,view_num=2)
```

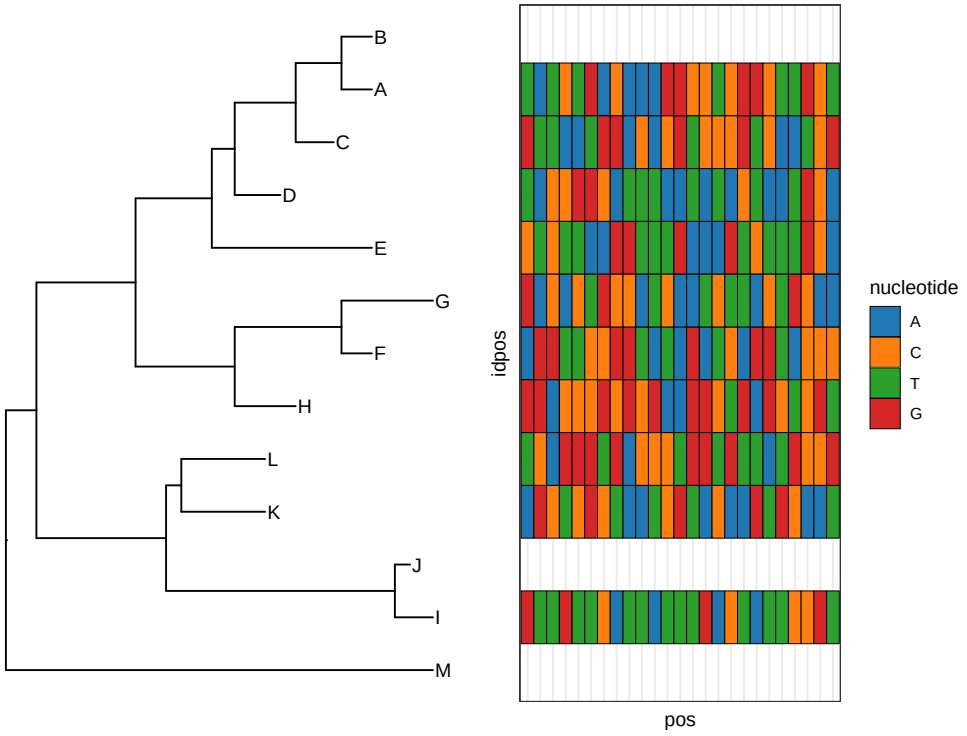



```
plot_view(component_specs,view_num=3)
```

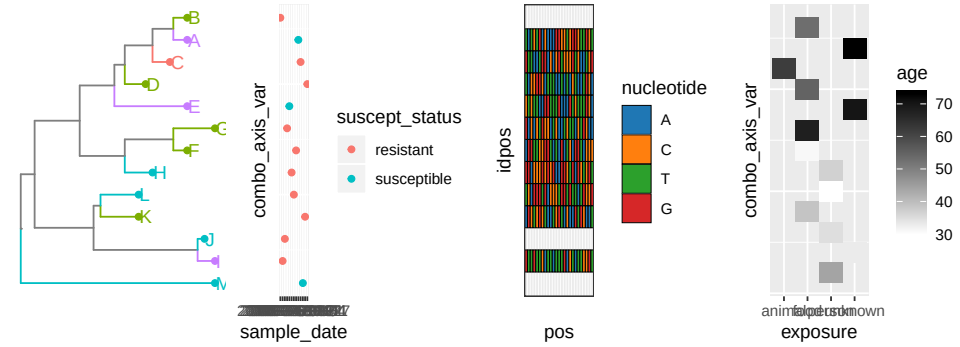


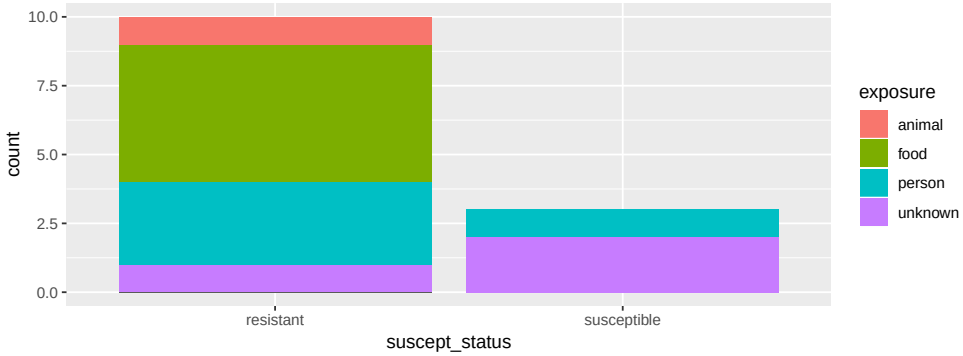



```
plot_view(component_specs,view_num=4)
```




```
plot_view(component_specs,view_num=5)
```


B
A
C
D
E
G
F
H
L
K
J
I
N
combo_axis_var
suscept_status
resistant
susceptible
sample_date
idpos
nucleotide
A
C
T
G
pos
combo_axis_var
age
70
60
50
40
30
exposure


10.0
7.5
5.0
2.5
0.0
count
resistant
susceptible
suscept_status
exposure
animal
food
person
unknown

Ebola Data GEViTRec

```
start_time <- Sys.time()
set.seed(416)

library(dplyr)
library(ggplot2)
library(igraph)
library(ggraph)

# 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 minimun to be in the same directory
#to add metadata to the shape file, you can also add .dbf files
gin_file<-"gin_admbnda_adm1_ocha_itos.shp"
lbr_file<-"lbr_admbnda_adm1_ocha.shp"
sle_file<-"sle_admbnda_adm1_1m_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")
```

VISUALIZING THE DATA

Modify the spatial objections to that they're more interesting to work with

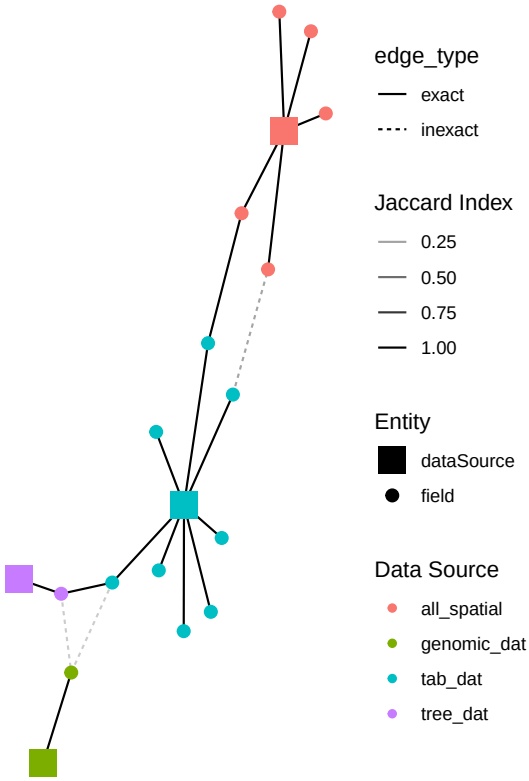
```
#join the spatial files
all_spatial<-join_spatial_data(gin_shape_dat,
                               lbr_shape_dat,
                               sle_shape_dat,
                               obj_names = c("GIN","LBR","SLE"))

#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
```

Also simplify the genomic data so that it actually visible. While its possible to view all 18,000 positions, to make the point here I'll stick the smaller handful that researchers tend to come up with at the end of thier analysis.

Data Harmonization

```
harmon_obj<-data_harmonization(tab_dat,tree_dat,
                               genomic_dat,all_spatial)
```

```
#plotting the entity graph
view_entity_graph(harmon_obj[["entityGraph"]])
```



Generate Specifications

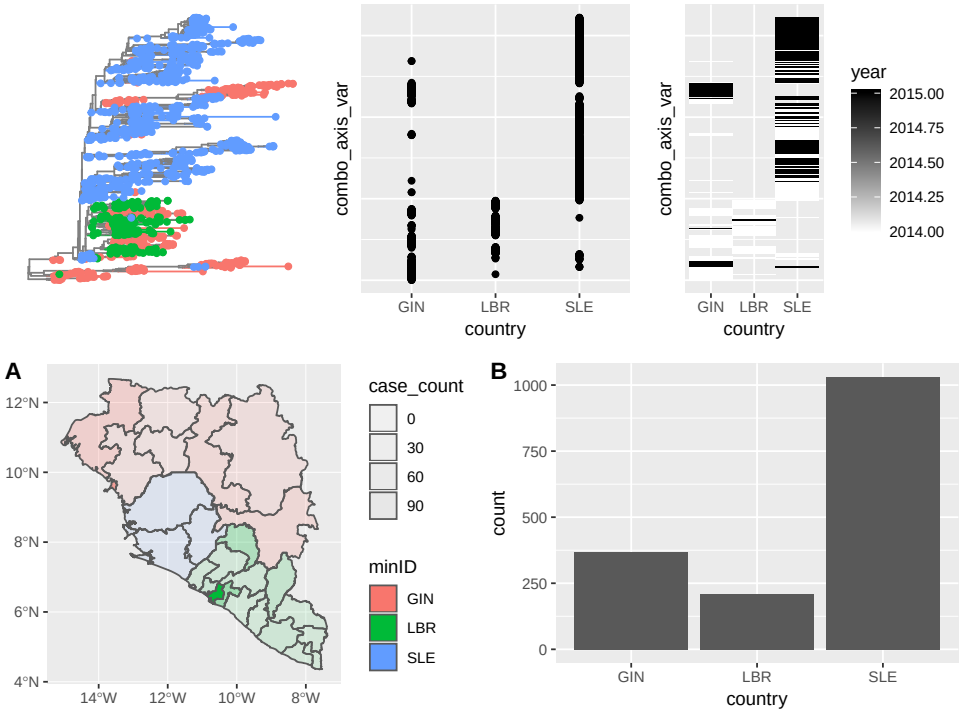
```
component_specs<-get_spec_list(harmon_obj)
```

```
## [1] "Generating possible specifications"
## [1] "Cleaning specifications up"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
```

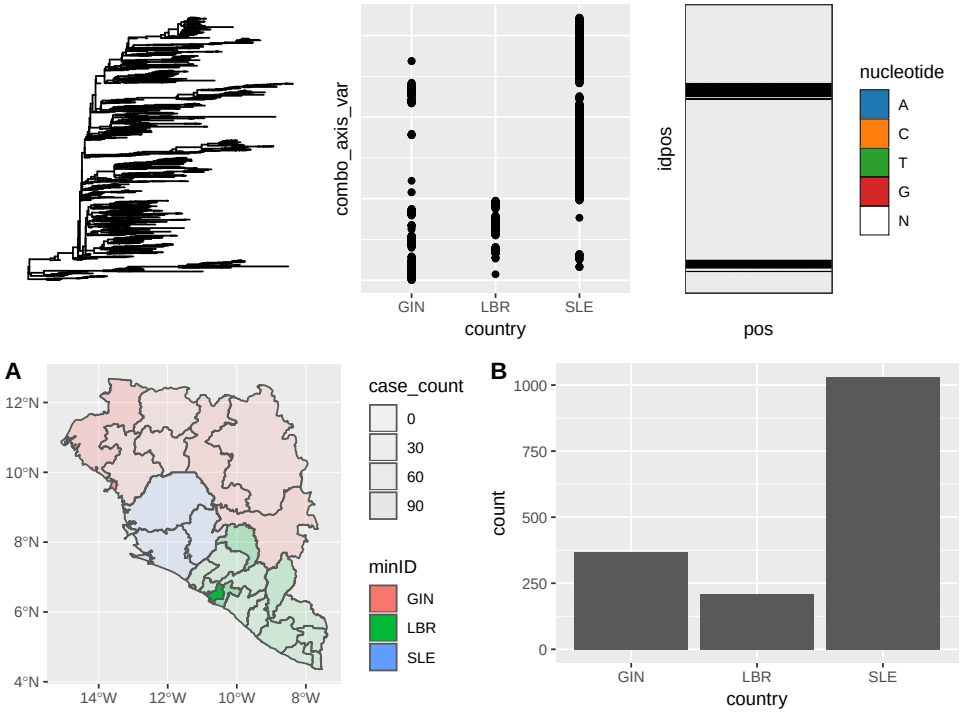
```
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
## [1] "Generating mincombinr spec"
```

Look at the generated views

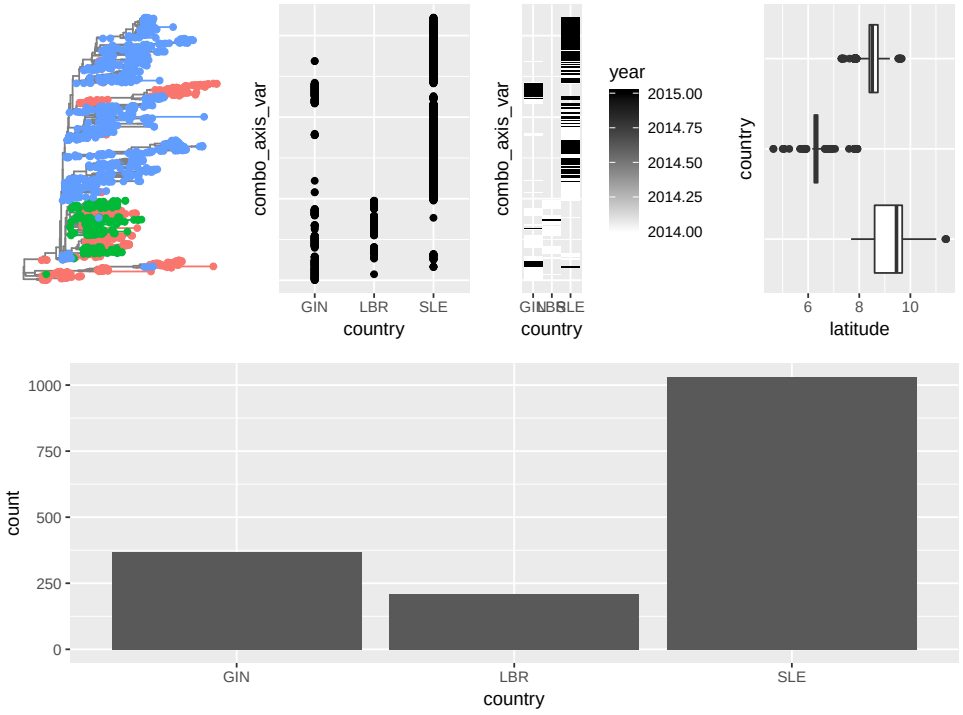
```r
plot_view(component_specs,view_num=1)
```



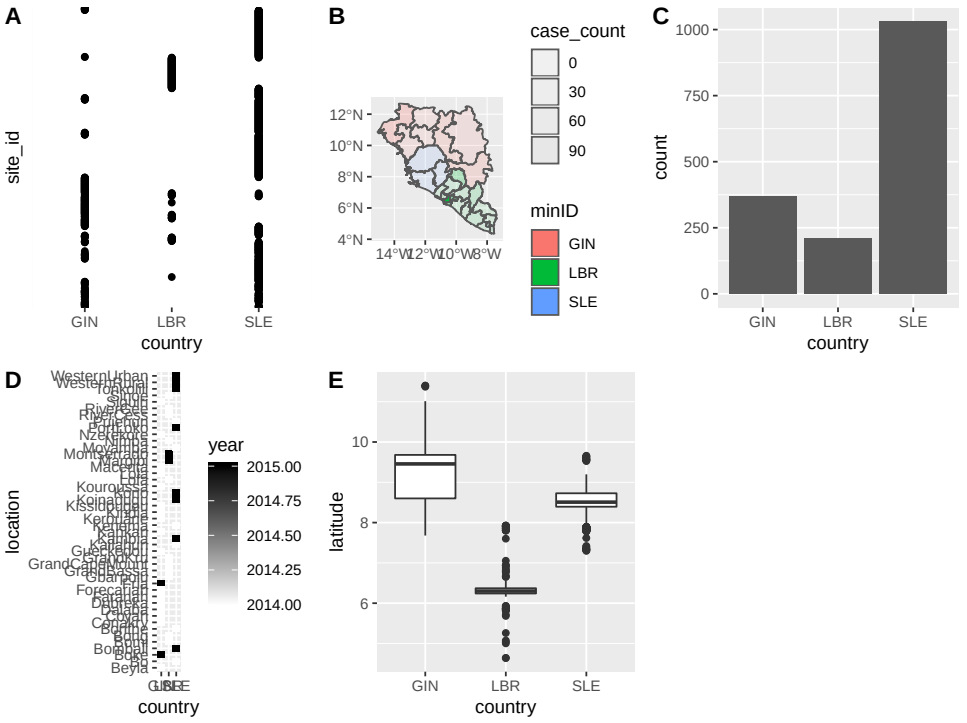

```
plot_view(component_specs,view_num=2)
```



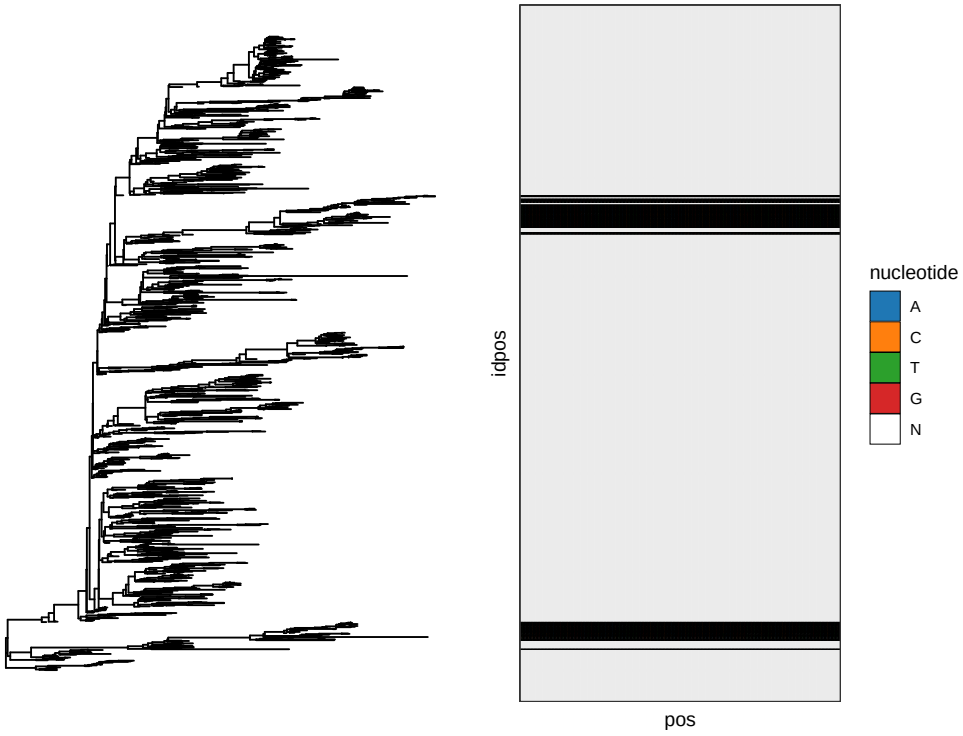

```
plot_view(component_specs,view_num=3)
```




```
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 18.04947 secs
```

S4 Survey Results

The following are anonymized results collected from the online `GEViTRec` Survey

S4.1 Participant Background

Q1: Please indicate your role in infectious disease diagnosis, treatment, management, and/or surveillance. You may select more than one option - Selected Choice

Role	Count
Bioinformatics - I primarily analyze genomic data	6
Clinical management - I work directly with patients, providing care and/or case management	1
Epidemiologist / Biostatistician - I have specialized knowledge and analyze public health data	4
Laboratory work - I work in a lab where I am involved with testing for different pathogens	2
Surveillance Analyses/ Data Science - I work with data to understand disease occurrence	5
Other (please specify)	4

Q1 - Other (please specify):

[1] "Research assistant - worked with patient data"
[2] "Business Intelligence Analyst"
[3] "Paediatric infectious disease doctor with PhD in gen epi"
[4] "Research Manager"

Q2: How would you rate your ability to program in R?

Experience	Count
Beginner - I program in R every now and then	5
Intermediate - I can confidently carry out analyses in R	4
Expert - I develop packages in R	1

Q3: How do you rate your ability of program in general?

Experience	Count
No ability - I have never used R and am not confident I could pick it up	0
Possible ability - I program in other languages and feel I could pick up R	0
No ability - I never write code	1
Beginner - I can write some code	4
Intermediate - I regularly write code for analyses or small applications	4
Expert - I can code complex applications that others use	1

Q4: How frequently do you produce data visualizations?

Visualization Usage	Count
Frequently - Use it to understand and communicate my results	7
Occasionally - Usually only for paper figures	3
Never or Rarely	0

Q5: Please indicate how often you use these difference data visualization tools. Use your judgement, responses don't have to be very rigid (i.e. if you don't use something literally every single day, but use it pretty often for your work, select the daily basis option)

option	I never use this	A few times a year	Monthly basis	Daily basis
Excel	2	3	1	4
Google Docs	4	2	2	2
Java Script (D3; Vega; Vega	8	2	0	0
Python (matplot; bokeh, seaborn, etc.)	6	3	1	0
R (ggplot, tidyverse, etc.)	0	4	1	5
Tableau	2	4	2	2

Q6: [Optional] What other data visualization systems or programming libraries that you use regularly (daily or weekly basis), but that are not listed in the previous question. Leave blank if you have no response.

[1] "Microsoft Visio" "Phandango" "PowerBI, Spotfire" [4] "BioNumerics"

S4.2 `GEViTRec` Assessment

Q7: Please answer the following questions about the session with

option	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
I could quickly find a useful data visualization from the set of suggestions provided by the system	7	3	0	0	0
It was easier for me to pick one of the suggestions than make a visualization on my own	7	1	2	0	0
It was easy to import data	8	2	0	0	0
It was easy to understand how the data connected together	5	5	0	0	0
Some of the things I thought of were not included in the suggested visualizations	2	2	4	1	1
The ordering of the suggestions made sense to me	3	7	0	0	0
The visualization suggestions included things I did not think of myself	6	1	2	0	1

Q8: Please answer the following questions on visualizations produced by `GEViTRec`

option	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
I would share these visualizations with others	6	3	1	0	0
The visualizations helped me understand the data	7	3	0	0	0
The visualizations that I saw were relevant to Genomic Epidemiology	9	1	0	0	0
The way the visualization suggestions changed according to the dataset seemed appropriate	4	5	0	1	0
These visualizations would help my current analyses	6	2	2	0	0

Q9: Please briefly describe the features of `GEViTRec` that you found to be most useful:

[1] "Good to have standardized API for data input; liked the harmonized dataset and the graph was cool to look at; liked the explore aspect that is automated and can lead to a surprise; good to have an analysis plan but good also to be surprised."
[2] "The way `GEViTRec` was able to provide quick recommendations based on the various types of data sets was user friendly."
[3] "I love how it does a good job of showing me the most relevant things. Algorithm is good!"
[4] "Linking of the datasets by key fields. Graph of the dataset linkage was quite useful. Initial visualizations for exploratory analysis."
[5] "I really liked the visualizations as starting points for creating my own visualizations."
[6] "I found it very usefull that it looks for connections between different datasources and shows you how they link together."
[7] "The entity graph was a nice addition, as it is something I had not thought to visualise. It would help to identify the best ways of linking data, if you were to go onto making other visulaisations separate to this. It looked generally easy to use and interpret, a manual of what format the input data should be in would be useful (or demo data)."
[8] "This tool is extremely useful for data exploration, particularly where there are incomplete, or highly varied types of data that must be integrated and displayed. Very useful for public health surveillance and initial review of data."
[9] "I thought the enity graph was very useful, and I liked the different combinations of visualizations"
[10] "Very quick way to visualize a lot of data. Automated linkage of disparate data sets."

Q10: Please briefly describe the features of `GEViTRec` that you found to be confusing or not helpful

[1] "Axis labels; size of some graphs"
[2] "Large data can overwhelm certain data visualizations with too many data points."
[3] "I find that with the hiccups associated with rendering the images I might be more likely to have it suggest a visualization, give me a broad sense of the structure, and then clean it up/visualize it myself in ggplot"
[4] "Amount of space given to each visualization suffered with large datasets. Tree and counts visualizations where too dense to understand. Did not provide a summary of over all findings."
[5] "Nothing was really confusing. Some of the suggestions were less useful which is to be expected."
[6] "I can't think of any, some of the visualizations might be confusing but you don't have to use them"
[7] "The choice of visualisations based on "relevance" depends on who you are looking and using the data, but as discussed this could be coutered by te fact that you may see something unexpected"

[8] "Some of the automatically-generated visualizations were not the most useful, although none were completely nonsensical."
[9] "I thought the program and visuals were all great and helpful."
[10] "Some plots weren't immediately clear as to what is being displayed."

S4.3 `GEViTRec` Comparison to Existing Systems

Q11: Have you used the Nextstrain or Microreact to visualize your own data or to explore some new dataset?

option	count
I do not know what these are	3
I have previously used Nextstrain and/or Microreact	7
I have not previously used Nextstrain and/or Microreact, but I am aware of them	0

Q12: How long do you think it would take you to create the data visualizations shown in the Nextstrain, Microreact, and `GEViTRec` systems?

option	A few days	A few hours or less	I can't generate this visualization
`GEViTRec`	1	8	1
Microreact	6	2	2
Nextstrain	5	0	5

Q13: Please state the extent to which you agree with the following features of data visualization systems:

option	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
Data visualization systems need to have interactive components	6	3	0	1	0
Data visualization systems should help me make visualizations	8	2	0	0	0
Data visualization systems should let me pick the visualizations I care most about	8	2	0	0	0
Data visualization systems should show the same visual representations even if the input datasets are different	0	2	1	6	1
Data visualizations systems should integrate directly with many different analysis tools	5	3	1	1	0

Q14: [OPTIONAL] Is there anything else you'd like to tell us about the comparison of

Nextstrain, Microreact, and GEViTRec that we have not asked you about? Leave blank if there's nothing else.

[1] "human made visualizations are fewer (3 vs. 5); machines speculatively create data visualizations, which could be useful (i.e. lead to surprised);layout and composition in human visualizations are more careful; need more interaction with maps in particular; needed to fix automated labels in GEViTRec"
[2] "I really like the capabilities of GEViTRec as a hypothesis generating tool."
[3] "no"

S4.4 GEViTRec Usage

Q15: Would use GEViTRec on your own datasets? - Selected Choice

Choice	Count
Yes	10
No	0

Q16: [OPTIONAL] Can you briefly describe the type of data you used or that you might use in the future? (i.e how many datasets, how big are the datasets, what types of data do they contain). Leave blank if you are sure that you would not use GEViTRec on your own data.

[1] "Antibactertial resistance bug, trying to look at relationships between metadata and genomic data, GEViTRec would be very useful to help with this quickly"
[2] "Potentially use GEViTRec on datasets with geographic information to identify outbreaks"
[3] "Genomic epidemiology, integrated with a ton of metadata!"
[4] "linked data; genomic data, link to patients, for enterics look to food history, current can do that and all the linking is done but needs to be visualized together; all the datasets are linked which makes it a good candidate; even though its all linked would be good to see how all are visualized together"
[5] "Combining genomic and epi data, especially exposure in foodborne outbreaks."
[6] "I'd love to be able to link geographic and temporal data together to make a nice visualization."
[7] "genomic, table metadata, geographic data, collection size 4000 isolates"
[8] "We have multiple data integration challenges. GEViTRec would be extremely useful for a range of different surveillance applications – particularly since it can be used to resolve genomic, gel, geospatial, tabular and other datasets, and can handle missing or incomplete data seemingly very well."
[9] "TB contact tracing dataset (n ~ 4000, 4 linked datasets including both epi and genomic data)"
[10] "Large patient datasets with lots of variables and some genomic data"

Q17: [OPTIONAL] If you used GEViTRec, please tell us about your experience using GEViTRec (i.e. did it help you understand your data or share it with colleagues). Leave blank if you have not used GEViTRec on your own data.

[1] NA "NONE"
[3] "Very helpful demonstration."

Q18: What do you think could be improved about GEViTRec ?

[1] "Already mentioned in discussions of system overview and survey questions."
[2] "Customization to user's background"

[3] “Consistency in aesthetics of the visualizations, maybe a bit more leading use of colour/shapes”
[4] “trees and data scaling; appearence of the chart and how they are all tied together, but would be good to overlay better. Unless told it would be harder to understand, but the linking is a strong point” [5] “The interface to set up an analysis could use a GUI interface.”
[6] “Letting the users rank their own preferences to tweak layouts, and perhaps preferences.”
[7] “interactive, ultimately web based”
[8] “I think it’s great!”
[9] “Refining the visualizations”
[10] “Perhaps a flag for missing data or errors in the dataset e.g. largely numerical column has some data points as letters”

Q19: Is there anything else you’d like to tell us about your experience using `GEViTRec` that we have not asked you about? Leave blank if there’s nothing else.

[1] “Nothing more to say”
[2] “None”
[3] NA
[4] “This can be really useful for genomic epidemiology.”
[5] “no”
[6] “no”
[7] “Nope!”
[8] “very intersting visualization tool”

S5 Survey Administrator Notes

A total of three out of ten participants were female. In our analysis, we did not find any results in the study that were due to the participant’s gender. We have thus made the decision to deliberately obscure the gender of the participants in these notes by using the pronoun ’they’ when referring to the responses of individual study participants.

S5.1 P01

During this participant’s session, there was an issue with the online survey platform. The administrator had to manually record the participant’s answers and add them to the questionnaire once the issues with were resolved. There were also some audio recording issues with the second half of the session. Overall, we relied on the administrator notes quite a bit for this participant’s results. Only the administrators notes pertain to spoken responses are shown here as the manually recorded survey responses appear in the the survey dataset.

- microreact, nextstrain, `GEViTRec` comparisons : “it’s like comparing apples to oranges, they do very different things”. Felt that minCombinR was much more for exploration whereas nextstrain and microreact were much more for communication.
- Liked the “element of surprise” that `GEViTRec` afforded.

S5.2 P02

- Interview was a little bit quiet because of the space we were in

- Less exposure to genomic data, although works in a public health role and has colleagues that do work with genomic data. Thus some visualizations didn't make total sense and they were not sure how to interpret the data
- Ebola example: "too many data points, it's hard to understand"
- Would like a slightly more tailored visualization according to profile. Prioritized some visual encodings and ignored the others.
- although stakeholder wasn't sure how to interpret phylogenetic tree, could start to see connections between the different graphs and asked questions.
- Found it interesting that "that you can just take data and with a few lines of code create a graph"
- Would like to use on "datasets with many variables"

S5.3 P03

- What stands out is that color coding is vertical rather than horizontal – heavily primed by examples. Looking for common color between rows.
- Looks a lot of at the x-axis labels
- Left to right reading
- Looking at everything together makes sense to them
- "It is telling me everything that I want to know"

S5.4 P04

- Scale issue : the phylo tree is too crowded, there is too much information present
- Tough to get close counts, still wants to be able to see more detailed information
- Geo graph shading is quite similar, it's hard to distinguish between the colours
- Last chart (bar chart) doesn't make add a whole lot of value for the screen real estate. Could just show numeric values and it would be just as good.
- Date legend is a little bit tough to read, could be standardized
- Density of data is difficult, everything is on top of each other. Could add a jitter to make it easier to see.
- Found that people read datavis down well. Would also like summary statements to make easier to read. Like the "report design" project.
- Idea for layout : show simplest information first, build the complexity in latter on.

S5.5 P05

- Finds EBOV also to be very busy, a lot of that is there. Likes the different sections of the visualizations and likes that there is a shared axis of information.
- In upper left, it's not easily obvious what the colours are, but does appear to correlate to country.
- Would prefer different chart for temporal data, that could show the data more effectively
- is useful, case counts are good to have in bar chart

- colour all look the same in the graph, like how it links
- overall, like the hypothesis generating aspects that machine approach might suggest.

S5.6 P06

- Said this year is "no more excel"
- In EBOV example: like the tree to be the biggest things, likes the combo axis. Encoding year data is difficult to read would have preferred a timeline. Likes the bard graph but wouldn't give it so much screen real estate. Doesn't quite understand the bar chart, although can see that SLE has the most cases. Finds some of the data a little be redundant. GIN seems to have a genetically distinct strain. I note that they are the first person that is drawn to the map first. Also found map shading hard to read (what's going on there?). Would like bar graph to be colour too because the black and white is note quite so nice. Would like to be be able to filter data, emphasize different things.
- Current workflow is to make 50% of chart in a program and then to use illustrator to get the rest of the way there.
- What make this useful to current workflow? -> Moving toward more complex data in the future and needs to see this data beyond a phylogenetic tree.
- Exploration of links is useful, want to be able to show data between geography and time, which is super important to them
- Finds the resulting visualizations that are produced to be much nicer than the black and white they get from their current tools
- good for "someone like me who is an advanced beginner". Said co-workers do often struggle with visualizing data and with vis tools.
- Big struggle is to assemble data, need help to connect the data a little bit more [I note that this can be done with other tools, again `GEViTRec` is a data viewer]
- Of the tools, thinks microreact is super easy to use and has used in the past.
- Extra views of the data that `GEViTRec` provides is helpful.
- "nice that it gives suggestions, but don now if each system *should* do that"; finds suggests are good to "break you out of your usual thinking".
- Would like mix : "I want it to look like something .. like here's a bunch of data go"
- struggle that a lot of fellow staff can't use R, so something that could lead to report would be very useful. Doesn't need to be quite push button, but something to help more. Was familiar with report design paper and wanted something like that.
- Also, their L.I.S system is rigid which can also make it hard to view their data more effectively

S5.7 P07

- Was interested in how relevance is defined, so broke it down for them. Would still like to see a deeper level of personalization.
- EBOV example: Look at the top row first, would like to see plots scaled for density of information. Finds that the map is pretty self explanatory. Overall find it fairly clear.
- Gives decent overview, but doesn't answer specific questions, for example wanting to "what are the age groups of SLE that got disease?". But still wants interplay between automatically generated vis and specific vis.

- Admits it's still helpful to see the things you're not asking the question for.
- Nextstrain and Microreact are very specific for some task.
- Usefulness of different tools depends on what you want to do. They are beautiful, but still not very broad in applicability.
- would like to further break down genomic data into clades, wants to be able to show different types of trees.
- Like machine generated solutions because "humans have biases to not think about certain things". "you never think to plot all your data together".
- doesn't code, but can do R and finds there is growing support for it. Overall, learning to code more is a non-starter because it's not actually use to their job.
- some code OK, but the simpler the code the better.
- "R is becoming more widely accessible and useable". So R is manageable. Extensive coding is beyond the scope of most people that they work with and themselves.
- "interactive is help. look away to something else. Interactivity can make future tool viability more useful."
- When you know what you want it should be clear and unambiguous how to achieve it

S5.8 P08

- EBOV Example: Main challenge is the dash board is there appears to be a color palette conflict
- would like to see tangle gram to connect tree with image.
- Getting multiple views with corroborating info is useful. Interesting.
- one of the few people that like time represented as a heatmap.
- like country level heatmap. But like others noticed that there was a problem with shading, where it wasn't super easy to see the differences in the gradients because the shading appears to be too similar.
- liked some of the redundancy, liked that the layout was consistent between different datasets (even though views changed)
- top three show similar data in different ways, give multiple perspectives on the data.
- it's an easy to use format, because you don't have to go back to tree and tease out the information that you care about. Can see it all there.
- also liked that machine generated views "it's easy to be lazy".
- run microreact and nextstrain servers locally, and it's quite easy to run
- finds human generated views are still quite tailored to viral phylogenetics. Seems a lot of potential in the `GEViTRec` visualizations.
- Advantage of having machine vs human vis : you can pick the obvious, but the machine generates views that can make sense and that you wouldn't have considered looking at.
- still need some curation that filters of what views doesn't make sense, but there's a need to balance human intention and machine derived results.

S5.9 P09

- "that's very cool". Typically just uses R to clean and analyze data.
- really impress by the speed, loaded and visualized data very quickly.
- EBOV example: because there is so much info, it can be overwhelming and hard to follow. All the colors can be a bit distracting, switching between black and white as well as colors. Feels that the issue is both that there is a lot ogf data and also there are a lot of visualizations. Feels that the visualizations make sense, but that it takes some time to orient themselves. Wanted to know how clean the data had to be in order to be loaded, I indicated that `GEViTRec` could actually be used to help clean data - they thought that was good use case and made it that much more useful. Found entity graph to be very useful, "nice to see how all of the data is connected together."
- Liked that `GEViTRec` gave you many options to view your data. Provides a different perspective.
- Thinks `GEViTRec` is a good base, but there needs to be more work to make the results more easy to consume.

S5.10 P10

- Has actually developed datavis tools for public health gen epi.
- EBOV example, it is obvious that most cases are in SLE
- Overall, felt that the visualization was useful, but was trying to familiarize themselves with all of the data. Felt overwhelmed by the scale of it.
- Found it to be pretty cool.
- fits within epi idea, usually you want to know "When, how much, and where"
- can see high level and also able to dig better in "how". Even when you have such data on hand, its still difficult to see it all together
- Glad that timeline can change: "hard to have a one size fits all, because outbreaks for different disease operate on different timescales"
- Found that `GEViTRec` was quite quick, really nice for exploratory things. Especially good when you don't know what is in the data"
- "Nexstrain is really slick, but there is a set thing that the are trying to do."
- "Overall, microreact is a traditional epiTool"
- Nexstrain show transmission, which is you have such data is great but not necessary
- "`GEViTRec` is like, here's your data, what in in, and if you want to carry on and build a transmission network you can make that decision".
- Would like a nice sleek interface to lets you explore the data visualization more easily.

S6 Survey Session Partial Transcripts

We recorded the full audio for each session including the study administrator providing the demonstration. Here we provide the full transcriptions for instances when the study participant spoke, either to ask a question or when prompted to do so by the study administrator. We include the time stamps within the session for the participants quotes and we also provided a qualitative code to indicate the quotation context.

Elided from these transcripts are the (rather repetitive) instances of the study administrator explaining the interface.

To contextualize these transcripts, we developed 10 codes through an open coding process:

- **Background:** Additional details on participant professional or programming background, including how they visualize data
- **Data:** Questions pertaining to the synthetic or ebola datasets, for example, where the data comes from or how it was generated. Also includes comments about data quality
- **EBOV Example:** participant responses to administrator prompting to interpret the EBOV data during the study session.
- **Entity Graph:** comments pertaining specifically to the entity graph
- **Functionality:** A functional component of the `GEViTRec` API. For example, the image annotation application, the syntax, etc.
- **Layout:** Participant comments about `GEViTRec`'s presentation of results.
- **Question:** Clarification about online questionnaire.
- **Scale:** Specific participant questions or comments pertaining to the scale of the data.
- **Tool Comparison:** Participant comments comparing `GEViTRec`'s functionality to Nextstrain and Microreact. Also comparison to any other tools that the participant brought up but that the questionnaire did not specifically query them about.
- **Usage:** Comments on how participant would use `GEViTRec` in the future and on their own data

S6.1 P01

Session Time Stamp: 03:21.51 – 03:25.91
Coding: Functionality, Image Annotation

```
01: Can I ask Question:

Admin: Yes

01: So I see X and Y and I also see y max and y max, what are those?
```

Session Time Stamp: 06:02.02 – 06:08.45
Coding: Functionality

```
01: The view method, is that a generic R method or \SYSTEMNAME\?

Admin: It is a generic method
```

Session Time Stamp: 17:39.92 – 18:17.92
Coding: Entity Graph

```
01: Can I ask a question. Do you have any functionality that can label the boxes or the points [in
    the entity graph]

Admin: Yes, I can do that but I haven't implemented that yet. That is a quick thing that I can do.

01: Cause I would have a tough time understanding, like, which data set is the purple data set.
```

Admin: Ah, so here we have the tree data set (gesturing the axis label), but what I don't have are the edge things. We could achieve that with some interactivity to make the entity graph more useful.

Session Time Stamp: 17:39.92 – 18:17.92
Coding: EBOV Example

01: Well it seems that I notice that on the tree here this green cluster, this green clade is matching up with the Liberian column I guess in this second plot there.

Admin: Yup

01: there seem to be some, like I can also see that this clade here with the red points is, well, seems to lining with with, is this Guinea ?

Admin: That's right, yes

01: Ok, so there's point there that are matching up both, here and here. guess, uh, I am just trying to get a sense of umm..

01: I guess one thing I am not totally clear on is what is being presented in these two graphs here in the top row on the centre and on the right side. .. [not audible..] .. the overall pattern is the data set to be quite similar and I do some date information on the type right. I see that they are both labelled with combo_axis_var but I am not totally sure what that is representing.

Admin: Right

01: Umm, well, I am also, I mean, having a bit of difficulty distinguishing the bottom left and the map diagram, there's boxes indicated case count and there are four boxes with different numbers but the colour seems the same. But I do see a variation in intensity of colours in the map, so it seems that there is one particular state or province or region that had a higher number of cases than others.

Admin: Do you know what country that case is in?

01: Umm, my geography is not the strong. Oh, but I guess by colour [pointing to the legend] it would be Liberia.

Admin: Yes

01: Umm, and, I guess one thing.. so looking at the map, looking at the overall,.. if the.. if the intensity of the map represents overall cases I do see one region of saturated colour would indicate a high number of cases, but if I compare that to the graph on the lower right I see that the overall number of cases actually seems to be highest in Sierra Leone. It has a larger [geographic] area so I guess that would mean that cases are more dispersed.

Session Time Stamp: 23:54.93 – 24:04.45
Coding: Layout

01: So on this one, do these point line up with, wait, the latitude?

Admin: Whenever you have this A,B,C they don't, but whenever you have the ``combo_axis_var'' that do.

01: I see

S6.2 P02

Session Time Stamp: 10:47.37 – 11:14.30
Coding: Functionality

[Showing the data once it's loaded]

02: Uh, would you mind if I just take a quick look what these look like in the environment? So if you load.. so for example the Ebola examples, when you load all the files I just want to see what the current state is..

Admin: [Shows demo of the data in the R environment]

Session Time Stamp: 14:38.44 – 17:23.98
Coding: EBOV Example

02: Well. I. first of all I just want to say that my background in genomics is not very strong, so I think, from my understanding, I think that it does make quite a bit of sense.. like.. just based on my expertise level. Um ... I think that ... I mean like.. this.. I think.. so .. if you have a lot of data points here this all starts to get very hard to understand. Umm ... an ...sorry just a quick question, what exactly are, so these are the specific countries here?

Admin: Yeah, and here they've got a combine axis, so what this is that is is that each of the points here you can read across..

02: Right yeah

Admin: So what it's doing is here is showing you a the distribution on the tree of the countries

02: Um hmm. Ok ... so.. well I think.. you might want to disregard my ideas about basically this part [point to tree] I just don't have a good knowledge background about the genomics. But this [points to tree and bar chart] is very useful information, the geographic map and also the case counts.

Admin: That's really useful to know

02: And that's just coming from someone that has more of a I guess.. more of a beginners knowledge of genomics ... so regardless of my background in genomics, this is represent very interesting information.

Admin: What do you think is interesting about it

02: Umm.. just that fact that ... umm how you're able to just take this data and with a few lines of code be able to create these sorts of graphs.

Session Time Stamp: 17:42.09 – 18:28.81
Coding: EBOV Example

Admin: So you have other information here that you would work with. Maybe you could talk to a microbiologist that would help you understand the information in this top row [phylogenetic tree].

02: yeah umm, I am sure this is actually very important information. I that a lot of times when you 're working with datasets with so many variables, just being able to have a code that lines up all these variables and matches them up on their own is very useful. Like I think it was a lot easier to understand when we were using the dummy data because there's way less data points and it's a lot more obvious to see.

Session Time Stamp: 24:03.74 – 24:29.46
Coding: Tool Comparison

02: Umm.. so like, if I were to use this one, like this program, like how long would it take me to do this? Or are you asking ...

Admin: If you had to make these on your own, without the help of these programs [how long would it take you].

Session Time Stamp: 24:54.51 – 25:17.84
Coding: Background

Admin: We are testing people at different skill levels. Is the tool useful for someone that's learn , someone that that's an expert -- that's what we're trying to figure out in this study.

02: Um hmm, and keep in mind that I also have no exposure to link either of these, whereas I have use R as well, so something like this [\SYSTEMNAME\] could be more easily produced like these trees [points to nexstrain and microreact interfaces]

Session Time Stamp: 28:00.43 – 28:18.66
Coding: Participant Comment

02: So I again, I guess like, perhaps, this program not geared towards who don't specifically work with genomics

Admin: Could I suggest that maybe you want even further customization based upon the user's background and preferences?

02: Yeah, yeah..

S6.3 P03

Session Time Stamp: 21:08.68 – 24:34.48
Coding: EBOV Example

[EBOV Example]

03: Ok, umm, so, I mean the fist thing that really sort of stands out to me is that, the colour coding seems to be like vertical rather than horizontal and just like what I saw with the previous examples I was looking for that sort of like key, not necessarily with the same rows, not necessarily between rows if that makes sense.

Admin: Yup

03: I am not sure if that's just a layout thing or not, but that's like the first thing that really stood out to me. Looking at the bar plot in the bottom right corner and thinking...ok.. not color coded...associated maybe.. Then I see the, the the different x labels and maybe I am just a bit...actually with all the plots the have the same x-axis. SO just.. yeah.

Admin: Ok, yup.

03: Yeah, that's just the first thing that I notice. Umm ... I mean .. I don't know if just because I am not entirely familiar in the top right, is that just showing the number of incidences in a given year?

Admin: Yup yeah. So in this case its actually.. if you.. umm.., what it does is is its a spatial alignment so you can go across here [traces finger horizontal], so all of these occurred in 2015.

03: And it's also telling me that most of it was in SLE. Yeah, ok, cool. Uhh, yeah I think maybe that's why the color thing trip me up, I was wanting to go from left to right.. umm.. when looking at things, and maybe just kind of got lost because the color thing wasn't doing what my context and intuition wanted me to do.

Admin: That's a really useful comment, thanks

03:Yeah, umm ... but now that I am looking at three of them together, I guess that it makes a lot of sense. I guess... yea...I don't know what else to say. I think that ... the more I look at it, I mean it's telling me everything I want to know, like how many cases, where were they, how do they cluster, in time, how do they cluster genetically, and like, what geographically, I see all of these things, so it's like what I would want to know.

Session Time Stamp:25:26.10 – 25:49.62
Coding: EBOV Example

[Looking at a different view]

03: So the top right plot, is in place of, or in lieu of the nap, that we were begin show before?

Admin: Yeah

03: yeah I like this, that's so cool

Admin: It's good that it's one of the higher ranked ones. I see that maps and phylogenetic trees tend to be highly ranked by the algorithm.

03: Yeah, that's really good.

Session Time Stamp: 26:00.97 – 27:26.87
Coding: EBOV Example, Functionality

[Question, Functionality]

03: That's like, cool, the more granular level might be helpful for certain reason that you may be producing visualizations of data. Like, to have receipts and stuff and to better understand. Are things being ranked for .... like what philosophy of data vis are they be ranked on based on? Like communicative or data analysis?

Session Time Stamp: 32:37.71 – 33:19.61
Coding: Data, Question, Usage

03: I guess for the next section, I'm definitely like ...

Admin: This section here? [Indicating to GEViT Usage]

03: Oh, yeah, sorry, I guess I should look at your screen. Yeah, so for this section here.. ummm... like I can give like the broad answer, but for the more granular questions, I just like don't know because I don't have all the data for my project data yet

Admin: Oh, no, it doesn't have to be granular, so for example one person's answer was that they'd use this on a drug resistance genome with lots of metadata. Like...that's acceptable..

03: Ok.

S6.4 P04

Session Time Stamp: 00:05.04 – 00:23.91
Coding: Background

04: Ok so, my visualization that I've produced are normally to how how a new data base will be built, not so much the data itself. So I do build them all the time, but I am not the end user, I am always the person whose building it for the end user.

Session Time Stamp: 01:09.02 – 01:23.11
Coding: Background

04: So for data visualization, that's probably a lot.. we go a tone work a lot of work with SQL server and background ETL pushes but that's all we do.

Session Time Stamp: 04:01.06 – 04:16.48
Coding: Functionality

[For data structure]

Admin: This data is a way of storing a lot of information so that it can be shared and read by the algorithm in a consistent way.

04: This is your own own, data structure?

Admin: Yes this is my own data structure.

Session Time Stamp: 06:00.77 – 06:07.37
Coding: Functionality

[For image annotation]

04: So those are all, points and squares.

Admin: that's correct, and when you do this it automatically records the x-y position.

Session Time Stamp: 06:45.26 – 07:12.67
Coding: Functionality

04: Who would be the large to annotate these images?

Admin: Often when we're investigating hospital or community outbreaks and you're actually shown a blue print of the area because you want to see how the disease spread through different wards. People often try to use an image for that. And especially for hospital outbreaks in particular, and one of the more challenge things is adding metadata to the image in order to make it useful.

Session Time Stamp: 07:56.37 – 08:29.17
Coding: Functionality

04: So that jpeg, will, if it's given to other people without this additional metadata, it's lost. There's now way to store it with the file?

Admin: In the current way that magick library works in R, the answer to that question is no. However, you can save the R data object and what I have done here, is I have actually saved those data objects, because they are just data frames and I've loaded them. So you can save the annotated data but unfortunately they [the image and annotations] have to be stored as two separate files.

Session Time Stamp: 10:15.68 – 10:27.72
Coding: Entity Graph

04: Oh, so you're saying that these four points all link the datasets together?

[points to entity graph on screen]

Admin: That's right.

Session Time Stamp: 12:04.14 – 12:27.09
Coding: Entity Graph

Admin: ... uses the graph to create recommendations for data visualizations

04: How does it deal with like, numeric values that aren't actually.. like.. linking fields.. they'd be like your 4th does of something? They're more `categorical numerical values?'

Admin: Right... at this time... we don't try to make such interpretations of numerical data. It would need additional information to figure out the context like that.

Session Time Stamp: 12:52.71 – 13:54.88
Coding: Entity Graph, Functionality

04: I guess ages would be the same, a person whose 30 years old.. well age is just a categorical variable?

Admin: In this case, the program would treat it as numeric variable and not a categorical variable. Any other kind of contextual information, it currently does not incorporate, although we recognize that as a limitation of the system.

04: And there's now way to change.. I guess you could build a way to change data type?

Admin : From the way that system automatically recognizes these variables?

04: Yeah, cause you have these quantitative vs qualitative, can you change those.

Admin: I haven't tried to do it, but if just look at the tabular data, you could just access the table data slot and you can do it manually. But there is no current interactive way of doing it that I have implemented.

Session Time Stamp: 15:47.76 – 16:19.78 Coding: EBOV Example, Layout

04: So you need to know that, you can read this horizontally

Admin: Yeah, I am trying to use the combo axis var to indicate that you read these graphs across. This one does not have it, so this just susceptibility status so this one is not read horizontally with the previous ones.

04: And your green matches back to here? [points between boxplot and phylogenetic tree]

Admin: Yes.

Session Time Stamp: 19:14.64 – 19:28.11
Coding: Data, EBOV Example

04: At what level are these [spatial polygons for Ebola data]. Like country level, or breakdowns like community levels?

Admin: there's country level, and within these countries there are second level administrative districts.

Session Time Stamp: 21:19.59 – 21:36.70
Coding: EBOV Example, Entity Graph

04: I see that some of the second level ones matches as well?

Admin: yes, there are some second level matches that is missing in here, or there are different spelling of things.

Session Time Stamp: 22:27.04 – 26:13.55
Coding: EBOV Example

04: So the phylo tree is way too small to really see where the all the sources come from. And so, these are your countries... so it's tough to get a total number of case counts for each one

Admin: For each one for what?

04: Well, however you want to stratify it. This ones seems to be countries so you can't quickly tell, I guess you have to add them up manually. It hard to see from the top view how many overall there were in there. The geo.. the graph of the different second level geography variables, the shading is quite similar, I can't tell ifs its a 0, 30, 60, or 90. I guess I can see the 90 is strong dot, but it's hard to see the different between the 30 and 60. The countries, the different shades of each country is quite easy. The last bar chart, doesn't add a ton of value, I think those could be... a numeric representation would almost be easy, where you could put a different visualization in there. This legend ... 2014.25 does that mean April? And .50 is June? I think those could be standardized.. unless that's the common way of interpreting this.. but I would have see this either as quarters or a brief month value. Am I interpreting it, the phylogenetic tree.. so this dot, it should represent.. well./. I can't tell how many dots are overlaid on top of there. I can tell from the phylo tree if it a big one or a small one.. but it's hard they're all ontop of each other.. so I guess you'd need to jitter it see how much data there was. Because I find with the phylo tree you can't tell size very well.

Session Time Stamp: 27:05.67 – 27:15.68
Coding: Data, EBOV Example, Scale

Admin: The phylogenetic tree is something the genEpi community really uses.

04: yes, it doesn't give enough real estate to it

Session Time Stamp: 33:59.91 – 35:43.45
Coding: Tool Comparison

[Nexstrain Mainly, some Microreact]

04: This one uses the sizes of the shape quite usefully, so you see differences between a small case vs a large concentration. Even though they do overlap there's a nice transparency that you can determine where the two are crossing.. well.. what do those colours represent, they must be genetic..

Admin: In this case... yeah

04: where the strains.. but the size are quite useful.

Admin: So this one [micoreact] has no size so they're right on top of each other

04: yeah so this one, is not as useful. And its.. this one.. I guess the lines represent.. what do the lines represent?

Admin: In this case it's transmission. So it hops from one area to another, and if we go to website , if you're curious, this is also an animation.

04: yeah i saw that. So you could probably.. see it over time. With those ones that go over time, do they also, do they just keep getting bigger and bigger and bigger, I find sometimes they also have to shrink down. Cause with Ebola, once you have the count and it gets bigger and bigger and bigger, but the amount of infected people does eventually start to decrease.

Admin: I don't know. Let's see [goes to nextstrain website, plays animation]

04: It does change. That's really neat.

Session Time Stamp: 36:02.12 – 36:45.11
Coding: Tool Comparison

[On how long it would take to produce these visualizations on his own]

04: Once the data is fully prepped and you understand the data?

Admin: yeah, how long do you think it would take to make something like this?

04: yeah.. uhh.. I don't have a tool that could produce.. I don't know... probably quite long. This one would not take long at all [\SYSTEMNAME\] this is a quite basic one. This one [nextstrain] would take longer. The phylo trees.. it really depends if you have the data and a tool that could actually use it. \SYSTEMNAME\ is definitely the fastest, it did this almost instantaneously.

Session Time Stamp: 36:59.81 – 38:05.63
Coding: Data, Scale

[Tries to adapt to whatever data you have]

04: How would this one scale up.. the computing time you show for those cases was almost instantaneous. But if you were trying to do ten year chart of hep C or something where you have tens of thousands of cases. It would be more.. instead of genetic... sheer volume.

Admin: I haven't yet tried it on something so big. Part of this is contingent on what the machine can fit into RAM. If doesn't load it into RAM, then it doesn't load it. So I haven't tried it

with 100,000 cases on yet. What I've tried to get more about is several different types of data and connect it together. In that case it was hard to find a real dataset with that much data in it.

Session Time Stamp: 38:32.63 – 39:23.17
Coding: Question

04: Uhh... what's an example of an analysis tool

Admin: Ok lets say, dplyr in R or some special library that you might want to run. For example a tool like Tableau, you can put your data in but its much harder to analyze it. For R, there's a much clearer integration between analysis and visualization, like between getting your summary statistics and then being able to put that into ggplot.

04: I see. So you're looking at something that can do.. like... your full stack of analysis, from like after analysis, to transformation, exploratory, and then publication?

Admin: Yes.

Session Time Stamp: 39:30.99 – 40:29.77
Coding: Functionality

04: umm interactive meaning drag and drop?

Admin: even what you say with nexstrain, the animation, zooming in, filtering

04: Yeah..dynamic.. like drop down and filtering absolutely. You're phylo.. like.. do most of the more targeted analysis for genomic data on the phylo tree you can zoom in and out and highlight certain cases.

Admin: Yeah, If I took \SYSTEMNAME's code and put it into a shiny application you could do that.

04: yeah, cause the colors all change and I just wanted to highlight the blue for example.. if could filter just to those ones, and it would effect all your other charts as well?

Admin: umm.. yes you can do that, yes you can do with shiny, no you can not do that in the current implementation of \SYSTEMNAME.

Session Time Stamp: 41:05.22 – 41:34.08
Coding: Functionality, Question

04: Data visualization should show the same visual representation even if the dataset are different ?

Admin: Yeah, if you have measles, staph, Ebola data nextstrain and microreact will always show you the same three things. Whereas with \SYSTEMNAME\ between the synthetic datasets and the Ebola datasets, it tried to change things up depending on what data you gave it.

04: Right, so if you don't have genomic data, why would you show a phylo tree? Is that what you're saying.

Admin: Yeah

04: Oh yeah you need to have that.

Session Time Stamp: 42:06.31 – 42:12.94
Coding: Usage

Admin: Would you like to use this on your own datasets if I got this stable enough?

04:Uhh yeah, I think there's a lot of potential for sure!

Session Time Stamp: 42:34.81 – 43:34.02
Coding: Usage

04: So for linked datasets, we do a lot of it and its always.. what we've just done.. which I think is really cool but needs a lot more work.. is taking genomic data, matching it to patients, which we then can match to ...for enteric.. you can take a look at their food history, like what they've eaten in the last few weeks, we can current do that. So we have multiple datasets where we've done the linking.. so that's an easy step.. we basically still have different datasets. And then.. well.. all of the datasets are really linked, you've got genomic data, risk factors, what they've eaten, and their risk categories. So even though we have them all linked together, it would be nice to see how they all visualize together.

Session Time Stamp: 44:04.94 – 44:30.98
Coding: Functionality

04: The appearance of the chart and how they all are tied. I know you told me that they could all be read horizontally, but if you could somehow overlay that with somehow tie it together that they're all linked together, and if I didn't know that, it would be harder to interpret. But I think this a really strong component of \SYSTEMNAME.

Session Time Stamp: 45:18.72 – 46:03.04
Coding: Functionality, Layout

04: People tend to read data visualization... well from what I've seen, read them down quite well like if you're scrolling through a blog post or something. Well you could probably have some summary things like you had ``1,600 cases over five years in three different countries'' and these countries have ``this count, this distribution'' and they you could get into more complex stuff. If found that with this, you had the complex stuff first and the bar chart second. So may you want to take them from a higher level and then take them into the details second.

S6.5 P05

Session Time Stamp: 00:08.10 – 00:40.30
Coding: Background

Admin: So you've said you never really programmed in R, or, you've programmed in R a little bit?

05: you know, I mean, more on kind of tutorial things. So I've use python a lot, and essentially tried to communicate data visualizations to epi I find tools like Tableau or Spotfire to be more useful, because it allows them too look at the data and filter and move things around, so they don't need to have any programming knowledge or things like that.

Session Time Stamp: 13:02.90 – 13:19.22
Coding: Data, EBOV Example

Admin: In the entity graph, we can see there are some dashed lines here and that is because the um .. DNA data didn't have all the same data points as the other datasets.

05: So does that mean that.. umm.. some of the data points in the dataset did not have DNA data associated with them?

Admin: That's correct.

Session Time Stamp: 17:34.79 – 21:24.30
Coding: EBOV Example

05: Ok first, it's busy, so there's a lot of data in there. Umm...look at that .. that's neat.. umm ...so the one thing that I liked before I see again here, I like the way that the different sections of the visualization, whether they're a scatter plot type thing, vs the I don't what you call it on the far right, and the tree on the far left, I like the way they share axes. I think that really helps when trying to understand things. And...you know...looking at this right now...ok.. on the upper left, it is not obvious what the colors are, but I see down towards the bottom middle, I think those are the color that are correlating with the upper left , which also correlates with the country graph in the top centre. They're both kind of neat but they are little redundant, because I see the colors in the tree are showing the same data and the middle graph on the top, so I think those are redundant. But I think the temporal data on the far upper right is really important, but I think there may be better ways of showing that data than the type of visualization it's showing there, because it's hard to distinguish between the colours on the right with the background color for the graph. you know, I think.. that might be...let me think about this for a second. What might be more useful, on the phylogenetic tree, those color icons should possibly use something like sizing to show temporal data rather than the way that's its displayed here. Looking on the bottom left, that's useful, showing the colours that it's associate with and.. the case counts are that's good to have a bar chart showing the case counts...um...actually now that I am looking at this.. in the middle [legend of map], I don't really understand the 30, 60, 90, what the purpose of that is.. because it looks all the same colors to me. I have not idea what that data is for. What would be neat, is you can take the data on the far bottom right .. the case counts.. and incorporate them into the map...and you might have done that. Like I see [on the map] if the different shades of the color are due to the uh concentration of cases, uh, and its proportion to the color on map, that would be really useful and I think that's what its doing. So in the far right the case counts are...well actually it's no redundant, because you can the amplitude of the quantity of the case counts, but it doesn't have to be that huge.. this visualization [bar chart].. so for the amount of space it's taking up, it doesn't need to be that huge.

Session Time Stamp: 26:04.43 – 26:21.96
Coding: Tool Comparison

Admin: have you ever used microreact or nextstrain before or know what they are?

05: I am familiar with both of them, I haven't used them for my own data. But I've seen presentations of them and I've gone and looked at demo data on those websites.

Session Time Stamp: 28:48.36 – 29:17.11
Coding: Usage

Admin: Do you think you would this on your own datasets if \SYSTEMNAME\ was out of prototype phase?

05: umm...I think I would give it try. What I really liked about it, and I mentioned i briefly, I like all of the hypothesis generating aspects of it. It might suggest some visualization options which you might not have considered.

Session Time Stamp: 32:59.98 – 34:19.11
Coding: Functionality

05: Yeah, and the thing that I like that you're doing here is rather than making a tool and throwing it out there that you're actually taking the time to do this type of study and talk to people and get feedback, and hopefully you're talking to a diverse group of individual in getting this feedback. And the only.. well.. you know.. so many times these data visualization tools are geared towards bioinformaticians, and people that are familiar with code, like coding in in R or coding in Python. The more we can make these tools easier for epidemiologists to use, or epidemiologists that are quantitatively inclined, the less code the better, I still think that your interface in R is going to be a little bit too much for some of the epidemiologists to swallow and adopt. And I did make a comment about a GUI and I do think that some to that could be optimized in just a very simple HTML interface, you know, in pointing to the data and hitting go. And I think you'd get a greater adoption from some of the less computationally inclined epidemiologists.

S6.6 P06

Session Time Stamp: 12:25.72 – 12:45.43
Coding: Functionality

Admin: So we now run the command to get visualizations

06: Oh sorry, have you scored the relevant of what, what you've found, for the different visualization somehow?

Admin: Yeah, so the way that we do it right now is that relevance is established by how a visualization is used in the [GEViT] dataset.

Session Time Stamp:23:35.97 – 27:30.17
Coding: EBOV Example

06: Well I always like the tree to be the biggest thing, but that's just, the personal bias, so I .. would put that on top of everything. But they you have your kind of sides ways, you can combine the axis on the other graphs. I find uh... like encoding the year data with a black and white colour to be a little bit difficult to look at. I think year data, I would like more of as a timeline, which I think.. one of the online ones.. microreact... they do a timeline kind of thing. So that part I find hard to read. I like the heatmap of the country, and the bar graph of the counts.. umm... I wouldn't make the bar graph so big.. but umm I like that.. I don 't... quite understand the middle. Umm.. I guess you're trying to.. like guess I think you're showing the same data in two ways, but maybe the colour of the phylogenetic tree, I don't know if you need to dots beside there, showing the country, because you're already showing it through colour I guess. So.. yeah obviously, SLE has the most cases form the looks of this.. um .. the pink country GIN, looks like its a different strain, or just distantly related. Um... and LIB, is probably branched off of SLE, maybe.

Admin: Ok, anything else?

06: Not quite sure.. like.. on the country map.. umm.. I guess, I guess I would think when I think country or a heatmap or, or maybe colour would represent number of cases more than location, I guess, you're just shading.. uh.. these shades are maybe just a little hard to read, if that's what that's showing. Cause, like distinguishing the case count is a little bit tricky based on colour or opacity. Um.. that one is a little bit tricky. Otherwise.. and then I... then just for boringness, I would personally just colour the bar graph and and maybe all the black and white elements I would also make colour so that I would have the same colour throughout.

Session Time Stamp: 29:06.78 – 29:26.73
Coding: EBOV Example, Scale

Admin: All 13,00 genomic positions so can't quite see them all.

06: Um.. I guess i mean you could filter down that down more to something that you found interesting. Maybe there's one region in that means something more.

Admin: Yup. It's up to the user to do that.

Session Time Stamp: 30:29.64 – 32:40.31
Coding: Functionality

[After EBOV example]:

06: I really like this it's very nice. I am a bit of a I don't know if its cheating, but I always just make something and then I get 80\% of the may there and finish it in illustration. I mean it's not always.. often doing just kind of one off things. You know you could have a website that does that as a one off per say. But ... yeah... I think this would be really useful.

Admin: What do you think it would be the most useful for? Cause I wouldn't say this is currently Illustrator quality images. But, why do you think this would be useful to your current workflow .

06: Well.. umm.. I am looking for something to.... well so I work for [....] so we're moving away from MIRU to whole genome sequencing, we have a lot of data and we need to represent it in in a way that is beyond a phylogenetic tree. And you... know... a lot of it... I think there's a lot of exploration that could be there... between different regions or countries. Unfortunately we don't have really fine epidemiological data, but showing relationships between time and geographic location is very important. And just something a little bit nicer than just the black and white newick tree, that you kind of have to manually fix up in some program.

Session Time Stamp: 33:08.19 – 34:08.34
Coding: Functionality

06: I think for like someone ... someone like me... I'm... I would say mostly an advanced beginner in a lot of data visualization and that kind of thing. Just to get back to my co-workers struggling with.. say you want to make a bar chart in excel or something. A lot of times they struggle with ... well they know what they want... they struggle with how to assemble the data and a structure that would make that happen. I guess that some of this is equivalent. I know I want.. some sort of heatmap or something, but how I would get there I don't necessarily... well .. I would need some help to connect those dots a little bit.

Session Time Stamp:36:36.10 – 37:10.25
Coding: Tool Comparison

06: I use them. I've used microreact a little bit, I've seen a lot of next strain visualizations, I haven't actually sat down and see if I could actually use it. I know they have the GitHub.

Admin: I know, but it takes a bit to set up. In that case, just say that you've previously use them .

06: I've found microreact really easy, just throw in a tree and table.

admin: And away you go

06: Yeah. Also, that's what I was looking for, just the timeline and the tree.

Session Time Stamp:38:10.17 – 38:49.21
Coding: Participant Comment

Admin: Whatever microreact produces is all you need and you don't need the extra stuff [that \ SYSTEMNAME\ visualizations] or do you find that some of the ability to look at different things is helpful or not helpful.

06: Um.. I think the extra stuff is very helpful. You know, a simple little bar chart that's great. Um... I haven't figured out how to make like a geographical heatmap that looks nice on my own. So I guess you'd have to find the shape files and put them?

Admin: Yes you would.

Session Time Stamp:40:56.22 – 41:22.00
Coding: Functionality

06: umm.. could you like manipulate the tree so that.. like... ape [R package].. ike adds annotations. And then would it just accept like a raw newick, or would it accept some sort of manipulated tree?

Admin: Yes. you can do both.

Session Time Stamp:41:56.25 – 42:50.29
Coding: Functionality

06: I think .. umm.. I am just struggle with, ``should'' and ``if''. Umm... it's very nice that .. you know.. it gives suggestions. I don't know if every system should do that, depends I guess what sort of level you're at. It's nice to get suggestions so that it also breaks you out of your usual thinking.. you know I always make this sort of thing.... yeah.. ok.

Session Time Stamp:44:08.36 – 44:24.63
Coding: Functionality

[Improving \SYSTEMNAME\]

06: Umm, so what kind of output does it save? Does it save like PDFs or pictures?

Admin: It does whatever you're able to output in R.

Session Time Stamp:45:05.08 – 45:24.00
Coding: Functionality

06: And you...umm.. envision just releasing this as maybe just and R package, or giving people this sort of... hold their hands with the R notebook type thing?

Admin: We're going to release it in R, but this is still a prototype.

Session Time Stamp:47:53.73 – 48:23.98
Coding: Tool Comparison

06: From my point, I found using something like ape [R package] tricky when I was getting into it. You know, something as simple as adding a picture to the nodes of the tree. You know it just took me a while to figure out, and something like the traditional R vignettes are just blah - I hate them.

Session Time Stamp:49:14.12 – 49:50.90
Coding: Functionality, Tool Comparison

Admin: You know if you want some very very specific type of customization there's always some point when you do have to do it yourself.

06: Yeah I think, um.. umm... I think that I am often just mixed between.. I want it to look like something and here's a bunch of data.... like you might have different experience. Like you might want to be like, here's a bunch of data, give me something to make sense of this or you know.. I kind of sort of know what I want and just make it pretty. Those are maybe competing things.

Admin: That's a really interesting observation. You're not the first person to say that.

Session Time Stamp:50:08.74 – 51:17.02
Coding: Layout

06: And, I'm .. you know I am just sort of looking for something.. you know.. we're struggling here .. a lot of the staff don't know how to use R and all of these things. But.. um.. I'd like to make something where we could get all the data.. like something for report. So I think a lot about what our reports look like, or how we could make something like a reproducible example. Like on your TB report project, to have something that isn't quite like just press a button. But some kind of integration that would help you take a whole bunch of data sources, and not just give you the same, you know it would be be just a little bit more flexible. Yeah.. I am not quite sure what I am saying.

Session Time Stamp: 52:35.11 – 52:41.86
Coding: Functionality

Admin: Thank you for participating today. I really appreciate it.

06: Yeah, it's really cool I am really impressed! Thank you very much. And some really good inspiration if nothing else.

S6.7 P07

Session Time Stamp: 00:15.70 – 04:41.99
Coding: EBOV Example

07: Ok, so.. on first look, obviously I am just looking at the axes to see what we've got. So country.. and so these combine axes.. the middle one I am not sure what it is showing me. Is it just where the cases are from?

Admin: That's correct.

07: And the second one on the far right is the year and the time distribution in which these cases occurred. Um.. I suppose that the tree.. I don't know how or what kind of tree it was generated or any kind of scale. I don't know if that's intentional or I would have known that because I

would generated that tree in the first place myself and so would have know what I had done. If that makes sense. And then, the map, yeah, I think it's fairly self explanatory in terms of location, and cases, and you've got co-ordinates on there. And the bottom right is case by country so you can see the burden of disease and its related to to the other two. And, SLE being the most obvious, eaah the blue, which is match is done on the map. So I think it is fairly clear, whether it answer the questions that I necessarily had is a different matter. But what I can see is fairly easy to understand I think.

Admin: Can you just comment a little more on the this.. that point that you said whether it answers the question I had or not. Can you elaborate on that a little bit.

07: so I guess.. as you've said, as you said the purpose is to generate a lot of different visualizations, which you can then use as you want. But the question I am asking might be more .. I don't know.. ``What are the age groups of''.. assuming this data is included in your table .. ``What's the age group of all the people in SLE that got disease in 205'' might be my specific questions. And maybe that appears on your fourth choice or 5th choice of visualization and maybe it means that the data is incomplete, but it's down there [lower ranked] for some reason. So I suppose as the person whose usually using this data at the user end.. so now you can introduce a vaccine strategy, or how do we isolate people, so from that public health or individual health perspective, I might have some very specific question that I want to answer. As opposed to when I am the researcher thinking about what is genomic epi of this disease. I might have much broader question which are address by this earlier image. So it's about picking the right image about what it is you want to know.

Admin: I've had some people bring up these statements before. Where they felt that what would be useful is bringing in some of the ..umm.. the specific background or research aims of the person as well. In terms of having that be a ranking criteria. So knowing that you care about vaccine efficacy, so that lower level of detail, maybe you want to see certain things that are different vs your other role.

07: Yeah, I think that's ... well.. I guess that's why I was asking what ranks. And you mentioned relevance. But I guess it matters who you are seeking relevance to. So if it is the person whose trying to resolve the outbreak, they may have very different questions to the person treating disease, or in lab try to identify what these things are. So all these people in the genEpi process have slightly different backgrounds or questions and so the relevance is of those is very specific for that person. So yeah, having that as a criteria for what is your questions I think would be very helpful to target your visualizations. It doesn't mean... well .. there's the counter point that it's something helpful to see things that you are not asking the question for because you may not of thought is something that is very obvious when you see it, either in a figure, graph, or tree. It may not be that obvious that those ones are linked, or that some age group is effected, until you looked at the data and it's something you didn't realize before. So having said, you always want to answer your questions with your specific figure, but somethings its good to see it a little bit more broadly.

Session Time Stamp: 12:15.19 – 12:33.55
Coding: Tool Comparison

07: I have seen nextstrain been demonstrated and I've used and seen microreact being demonstration.

Session Time Stamp: 13:13.29 – 17:10.57
Coding: Tool Comparison

07: So when you say ``human created'', so microreact and nexstrain, do you mean that humans have decided what these three things will show me?

Admin: That's correct

07: So you decide to show a map with lines on. diversity graph.. yes.. ok. As opposed to you where the algorithm decides which is the most relevant.

Admin: especially thinking back to your question earlier about the specific research questions you have in mind.

07: Yeah I think this is it.. I thinks it's really specific. So I have used microreact in terms of loading data and looking to think ``oh that's kind of interesting''. But I was never very interested in using it because all my isolates are usually for from very specific places and phylogeography is not very interesting to what I am doing, so instantly one-third of that figure... of that visualization... is not that useful. But with next strain and microreact it depends on what you're doing and who you are and what you are trying to tell. So I have seen similar things demonstrated for both Zika and Ebola outbreaks. Very beautiful, you know, maps with figure bubbles in real-time, and you can play the next strain.. and to see how it's transmitted from one place. Which is really useful especially if you are trying to sell that to some government agency especially if you're trying to say ``look let's close the border'' or you know ``We need to move... isolate people in this region'' because look its unaffected but is probably going to come here next. So I can see how if different settings different ones [visualizations] are really useful, whereas, you're trying to break down the individual genomic clusters or clades that are responsible for disease and spread. The tree that they [next strain] gives you might not be the most useful tree. You might want to visualize your tree using a different program... or a different settings, or time based, I think it really depends on who you are and what you want to show. And form my perspective as a researcher, you have many different things that you want to see on a different day and with different questions. So you can make the figure or the graph based upon whatever thought comes into your head.. you know.. ``i want to see all the isolates from 2015 on a tree now.. go'' you just draw it. Its about where your hypothesis is. So the power of these ones [\SYSTEMNAME\] and making a machine produce things is that we haven't actually had the bias to not think about things. So, would I use it necessarily to draw my graph of age vs... I don't country.. possibly not.. because I know exactly what I want to do and I can just take the data and do it myself. But would it help you realise something that I hadn't thought about, that for example age is related to clade, that I hadn't thought of because you never plot ever single data point against every other single data point, because, I guess that's takes hours or days to do that. But if brought up something really obvious that has implications, like ``oh, okay, all of these are circulating among children or region of the country, why is this?'' well, then it raises questions that you would go on to answer. So I guess what I am trying to say is, that I suppose the power of the machine you haven't got a human bias. Which in some ways its good, because when you know exactly what you want to do you can bias it an make exactly the tree or figure you want, but when you don't know, when you're discovering things, then maybe it is good to have things that are generated without human input.

Session Time Stamp:18:11.26 – 20:36.75

Coding: Functionality, Tool Comparison

07: So just a point about how long it would take me to create these things. SO I think the other limiting factor of these visualization tools, well not just visualization tools even analysis tools, is that its always my issue is that I don't code. And I thought about learning to code ... and well A) It's not going to help my job as a doctor (coding) and it would take a lot of time and energy to do that. So finding things that do what I need to do simpler, using either website based or R, okay R does involve a bit of coding but not complex coding, I think is really powerful. And as genomic epi becomes more established in healthcare and other settings where you have non-specialists not only just trying to visualize the data and understand the data, but you know, slightly more specialists trying to analyze the data. Having everything based on command lines that most health professional can't approach is an issue.

Admin: Yeah, one of the things for future versions would be take this [alg] and make it into a more

friendly user interface. That's future work, right now we're just testing out the algorithm and what it produces to give us a sense of..

07: yeah and I think R is becoming something that is more widely accessible and usable partly because its free and just there's just so much online help to do things. You know I didn't know how many of these things, and being a relatively older person learn to code, as opposed to being a teenage back in the day, getting to use R in manageable and I think many people find that. But doing more extensive coding is I think beyond the scope of most people who are already professionals getting into genomic epi, who want to use the data and want to use it for public benefit or health benefit or whatever, that don't have to back to basics and have to work out how to analyze these things from scratch. I think the interactive bit that you mentioned earlier is also very helpful. Because I think that you want to see something and to see if it's interesting or not, and if not then you can just flip to something else without having to do a whole separate analysis. So I think that it makes the usability much much great.

Session Time Stamp: 21:10.20 – 21:33.08
Coding: Functionality, Tool Comparison

07: So is there, I've not come across anyone else that's done this and I haven't necessarily looked , but are there other forums where you can do a similar thing. Where you can just upload your data and say, you know, ``show me all your visualizations?''

Admin: Um.. the answer is yes and no. [Explains Tableau ShowMe, mentions some other research systems. Says that no, nothing is specific to genEpi like this system it].

Session Time Stamp: 22:29.14 – 22:44.00
Coding: Functionality, Tool Comparison

[Value of connecting data]

07: Yeah, I think that's what we don't have. So yes, draw your tree in one program and suddenly you want to annotate it with some data that's in another program and that's what's annoying I think at the moment.

Session Time Stamp: 23:42.64 – 24:19.29
Coding: Background, Functionality

07: Yeah I think as you say, it might be very much dependent on who you are and what forum your in. So in a very research based forum I think it's ok to be very exploratory. But of you're talking about using it in a much more kind of focused public health level or individual health level. Then your questions are very clear and the outputs should be very clear. And, I don't think you're going to.. yeah I think it just depends who is using your tool or who you are targeting your tool at.

Session Time Stamp: 25:17.67 – 25:34.49
Coding: Background

07: Its nice to see what is being done by the data visualization community. Because it is all very well for us to be drawing our graphs and trying to explain our data but I think we need to do a . .. we're not necessarily the people to know how to do that or know how to do that, so we might as well go to the people that do.

S6.8 P08

Session Time Stamp: 07:16.27 – 07:25.32
Coding: Functionality

[On the entity graph linage derivation]

08: Is it doing it by exact pattern matching by categorical variable or is there some fuzziness to it?

Admin: There is some fuzziness to it, and you are correct that we are using categorical variables. Your intuition is bang on.

Session Time Stamp: 12:24.18 – 12:35.08
Coding: Functionality

[On relevance ranking]

08: Relevant to whom?

Admin: Relevant to someone study genomic epidemiology in the general sense

08: Right, the obvious linkage that people might be interested in.

Admin: Yeah, but also the kinds of visualizations that they're going to want to see.

Session Time Stamp: 15:24.80 – 15:46.99
Coding: Functionality

08: Oh, that's really cool. I mean the only challenge here and giving this to you as a dashboard of potentially visualizations is umm I mean if someone is very visual you end up with a palette conflict, just because of the way the colours are used. But other than that, these are all extremely useful visualizations. And having essentially a menu to choose from is fantastic.

Session Time Stamp: 16:55.87 – 17:32.34
Coding: Functionality

08: You could almost do like a tangle gram type thing across the two. But I agree in order to do gel images.. I mean.. the gel annotator that you showed earlier is very slick and has a lot of potential. But in order to actually re-order stuff you would have to deal with gel normalization and deal with lane boundaries, and then you'd have to slice and dice and to reorder the gel image ... and I don't think you want to do that.

Session Time Stamp: 23:01.15 – 23:17.58
Coding: Data, Entity Graph

08: There's still a fair bit of missing data in the tree by the looks of it.

Admin: yeah, the DNA data is only available for GIN here.

Session Time Stamp: 24:34.26 – 24:50.56
Coding: Functionality

[On data quality]

08: Well, that's cool, how dependent is this on structure vocabulary, especially in the tabular data where a lot these linkages are made.. and does the lack of a consistent ontology... or.. so does that come through in a lot of the data cleaning steps up front?

Session Time Stamp: 27:18.65 – 31:47.98
Coding:EBOV Example

08: So I like it for one thing. I think that you are getting multiple views of the data and you're giving... probably.. for instances, in the first two panels where you've got the tree, and you've got essentially.. I don't know what you would call that chart, but you can see is the genomic breakdown by country, and i think the time as a heatmap is nice as well, if you're looking at the timing of the exposures and the history of disease. One question though, for the A) when you pulled in the various shape files and essentially had the custom code that basically group them, did that involve shape file relationship and docking the appropriately or docking them appropriately in two dimensional orientations?

Admin: This function I wrote called ``join spatial data'', which is not custom code but shipped with \SYSTEMNAME\ does do that.

08: Again, does it have those relationship built into the shape files?

Admin: no, the individual shape files just have the co-ordinates for where those polygons should appear, so as long as those shape files are loaded correctly as spatial data, its a very straightforward relationship to resolve because that's where the polygons belong. It's absolute and not relative. So this doesn't take a lot of work to resolve, because there is no reason why they would overlap since they contain relevant geographic data.

08: I like the heat map by country. All of this is a very useful dashboard if you are trying to take an initial cut of the data and trying to see where there are interesting things here.

Admin: ok

08: um.. you know one thing I guess, is in that panel A, well, these are all picky points, but the gradation in shading on the case counts is so low... you know there's only one place in LIB where you have an obvious hot spot, but outside of that you know, the shading levels don't look different enough to be noticeable.

Admin: yup, a lot of people have commented on that. Um.. I am going to ask you some pointed questions that I have heard from other people, which is stuff about lay out.. um.. what do you think of the layout? Could it be improved? Also redundancy of information. Some of it is redundant because there is not a lot of data out their publicly. However, those are two things that people have pointed it and I'd like to know your specific thoughts on it.

08: No, I mean.. like the demo example you gave before, its having a fair consistently layout, where the top three and more or less linked. And those top three panels, are giving you, similar data in different ways. But they also tease out different aspects of it that you might be interested it. For example you might be looking at the recency of cases or the geography distribution and panels two and three there let you see that very specifically without having to tease it out. I mean its obvious that you're getting the same information, but you're just getting it in different ways and its much easier to consume in these data. For me, if I were taking a first cut at surveillance data, I like having two different views because.. for one.. its very easy to be lazy and if you are trying to find a specific piece of information or answer a specific question um.. you know.. having a few different ways of looking at it rather than actually taking a highlighter to a graph and tease it apart is very nice.

Session Time Stamp: 32:35.10 – 32:57.15
Coding: EBOV Example, Scale

[On usefulness of genomic data visualization and scale of genomic data]

08: I don't know how useful that third panel it is, other than it shows you the paucity of data. For example, you know that you have only some of the sequences loaded, so you know, seeing where you had sequence relative to where you had the tree is useful...but..

Session Time Stamp: 34:36.83 – 34:50.23
Coding: EBOV Examplem Scale

[Another statement on phylogenetic tree genomic visual with data scale]

Admin: So without additional filtering all you really see is that yes this had data and this did not.

09: Yeah, i mean that's this is really useful for at this level of resolution.

Session Time Stamp: 38:54.21 – 42:43.41
Coding: Tool Comparison

08: Right, on this question, how long do you think it would take you to make these data visualizations.. you're talking about from scratch,

Admin: Yeah, I mean excluding, well I know nexstrain has a big data analysis pipeline attached to the back of it. So.. assuming that you've run that already and now you're just going to visualize the data.

08: Uhh yeah.. ok.. well I mean, we run both microreact and nextstrain servers locally.

Admin: So it's very straight forward then?

08: Yeah.

Admin: Just out loud as a question, what do you think of the difference between the human and machine generated views. Um.. I think there's a lot of potential. I mean, the human generated views are based on.. you know.. a lot of tailoring over time, and they are very geared toward viral phylogenetics. Which is nice, because you know [names] .. a lot of the work they're doing now is about applying bacterial or [name]'s work is focused on that. You know, the fact there there is, but you know, the advantage of having machine generated visualizations is that you know you can pick sort of obvious visualizations that people are going to be interested in, and I think those are fairly well hammered through. Umm.. but I think the nice thing about having some machine generated visualizations is that, it looks at, associations that may make sense, in this system but may not be a conventional way of visualization something. So you might get an unconventional view of the data that could be useful. So as the model gets tighter.. for example you point out the example of latitude that maybe is not the most useful visualization to have set up.. so you know.. there will be some back end curation at some point to filter out the the things that don't make sense, but you know, I think, sometimes having some of those weird data data views where you know, ``What happens if we actually do it based on this fields '', those are necessarily counter productive. You know a lot of them are just kind of fluff, but some can actually provide useful insight into the data.

Admin: Yeah, we're trying to move towards that, and to think about context to make that first view of the data even more relevant.

S6.9 P09

Session Time Stamp: 07:01.92 – 08:18.41
Coding: Participant Comment

Admin: \SYSTEMNAME\ tries to find the connections between datasets using categorical variables and then it actually analyzes the results graph of those connections to figure out what it should visualize and also how it should prioritize what visualizations it should show you by picking paths that could produce relevant data visualizations

09: That's really cool. Yeah, its like, that's so I use R in a very you know clean data, analyzing data, sort of basic analysis sort of way, But seeing it used in these kinds of contexts is, for lack of a better word is really really cool.

Admin: Yeah, the thing is once you get comfortable with R and work your way up you can actually do a lot of things like this, because this [generating data visualizations] is just another kind of analysis.

09: Yeah

Admin: And so part of the reason why I like using R is because a lot of that analysis tools are super relevant to a biostat and epi community and you can add these kinds of things on top of it. So you say ``Oh I've done my analysis but now I want to see what kinds of visualizations I can make, so let me try loading \SYSTEMNAME\ and see what its does..'' that's the ultimate goal for how this would fit into somebody's work flow.

09: I like it, I like it a lot.

Session Time Stamp: 22:14.18 – 25:29.26
Coding: EBOV Example

09: So, first I was really impressed by the speed of it. I do know that's a bit of weird one, but it loaded that data and the visualization very quickly, so I was very impressed with that. In terms of the actual visualization, umm, because there's so much information in here, it's a bit hard to follow. It's a bit overwhelming I think. And I think because of all of the different like, the fact that some of the graphs are colour scales, and some are black and white, it again makes it a little hard to follow and it's a bit distracting. However, I do like the fact that its still, you can still read everything horizontally across. Its not as easy to process as the sample data was, which you could take one look at it and say, I totally understand everything that's going on. This one would take me a few more minutes of just sitting and thinking and needing to digests a little bit.

Admin: Do you think that because there's just soo much data, or because there are a lot of visualizations?

09: Umm... I think its a little bit of both. So, because there's a lot of visualizations and because there's a lot of data./

Admin: Now I am going to ask you a bit of pointed question here. Some people have said.. well... like you've said you don't have an in depth exposure to genepi but you've seen some stuff with it..

09: Yup.

Admin: So, one question I have, are there visualizations in this [EBOV View] that don't make sense to you, and, you're just thinking `I don't need to see it' or `I just need more time to grok it '

09: Ummm... no I think, I think they all make sense. Although, getting a little bit.. the bottom left one with the map, and the case counts, and the fact that they're all a gray scale doesn't really make any sense to me.

Admin: Now, if you were to take a stab at kind of making any kind of interpretation with this, what are some things that pop out at you, that as like as first impression you're like ``oh that ...''

09: So, right, do you mean in terms of, interpreting the data.

Admin: Yeah.

09: First stabs, case counts in SLE were the highest, umm, lots of cases in 2015, umm, kind of first impression it's kind of what I got.

Session Time Stamp: 9:40.63 – 31:49.25
Coding: Data, Entity Graph

09: I'm just going to make a quick comment while I am filling these in. What I am realizing with this program, that's it just like, most other programs you can't just expect to put in messy data and get beautiful visualizations. Right?

Admin: Yup

09: Like as long as you put the time and effort to make the data readable, for lack of a better term, it will come up with really great visualizations. But if you put in data that kind of.. that isn't quite as clean as it could be, your visualizations aren't going to be quite as clean as they could be.

Admin: Yup, you're totally right, so there's two components to that. Its a data viewer, it basically, if it's garbage in to some extent it's garbage out.

09: Yeah

Admin: So there's two kind of ways that the current implementation can work with that. So one is, the entity graph, even when I was putting in all of this data together, I had to clean it up. But what I found useful about the entity graph was when I thought there should be connections but there weren't. I thought `why isn't this work' and then I would take a look and see, and realised that the way guinea was abbreviated in the tree data and genomic data. And those connections I wasn't expecting didn't show up.

09: Yeah

Admin: So that's how I used the entity graph was to help me trouble shoot the data clean. The other reason this is in R is precisely because you probably have to clean up the data, and you can do that in your R code and run \SYSTEMNAME\ in your R code once you're done.

09: Totally! I am actually think that the entity graph, am I saying that right? I am saying that is incredibly useful as well. Like it's really nice to see that visualization of how all of these different datasets can connect together.

Session Time Stamp: 33:04.09 – 33:19.39
Coding: Data, Entity Graph

Admin: As I already said, I used the entity graph to clean up the data, which was not a task that it was designed for, but that it's useful for.

09: I think that is, yeah, I think that's a very useful approach to it.

Session Time Stamp: 35:38.36 – 35:48.31

Coding: Tool Comparison

[Showing map animations on the nexstrain web page]
09: That's really cool, but that's mainly because I really like animations

Session Time Stamp: 37:00.46 – 38:39.40
Coding: Participant Comment

09: Um, so , I think this is where, my potential lack of experience genomic epi data might come in. On all of these I am ``There is a lot of data''. it will take me a moment to process. Umm.. I ... okay... can you scroll down a little. Perfect OK. I just wanted to take a look at all three of them again.

Admin: These are also in your survey as well.

09: Ooh, ok.. ph that would make more sense, there we go. Ok um.. so yes, there is a lot of data, umm... I like how the \SYSTEMNAME\ allows you to kind of customize your views to what you would like... or... gives you a lot of different options for how you want to visualize things and you were saying the other two were a bit more fixed...?

Admin: Yeah they're fixed, that's why they've [the creators] have chosen [to show you] and that's pretty much all you see.

09: .. and I mean... great but maybe you want to visualize your data in a slightly different way. So I like the fact that \SYSTEMNAME\ allows for those differences to be incorporated. I will just say that from a visualization perspective, like from a visual perspective, umm.. nextstrain and micro react.. like... their colors work really well, I am not exactly sure what they are trying to tell me but the are visually appealing. Does that make sense?

Session Time Stamp: 38:55.85 – 39:18.00
Coding: Functionality

09: Yeah, like, it [\SYSTEMNAME\] I feel like it's got a really face there.. umm.. I don't know if they idea is that people will see these visualization and then they can tweak them to their linkings, like change the color scale, change the formatting of the graphs... of if that's supposed to be the finished product.

Session Time Stamp: 39:37.03 – 40:12.35
Coding: Functionality

Admin: We're getting feedback from others too that the first view is decent, but the visual, the aesthetic appeal, needs to be improved a bit. So that will a subsequent iteration.

09: yeah, and I think that is a, that is kind of exactly what I am trying to say. Where it's like ... the information is ...is great. And its' really interesting that you can use these different combinations and you're give give different options or however many different options , but it just needs to be tweaked or refined a little bit.

Session Time Stamp: 44:40.67 – 45:00.47
Coding: Background

Admin: Thanks for your time and thoughts.

09: Umm.. no problem, I.. like.. I think its really.. again for lack of a better word, I really wished I used more genomic data, I think is really interesting and I can see the potential use for this.

S7 Comparison to Previous Work

GEViTRec: Comparison to ShowMe, Voyager, and Draco

Jan. 2021

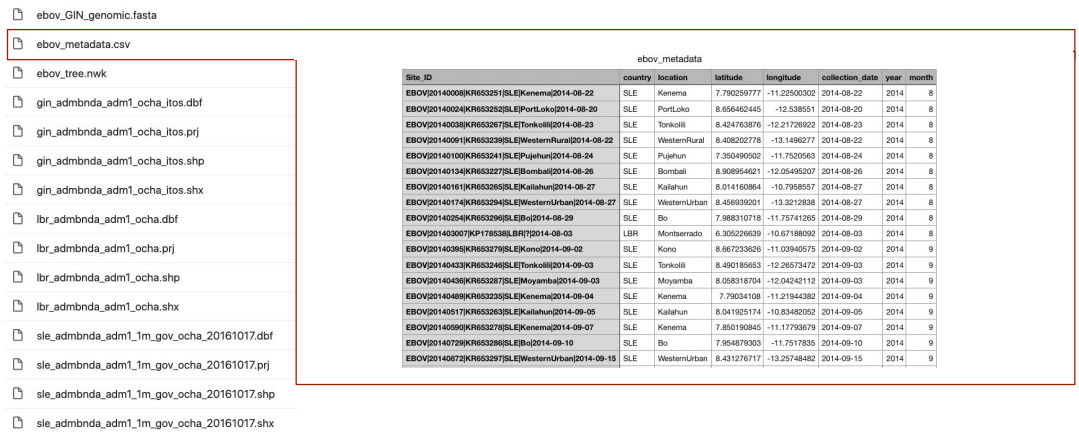
Comparison Details

There is no direct comparison between GEViTRec and other visualization recommender systems. Not only is GEViTRec informed by a VPDS that embeds domain specific visual design alternatives into the recommendation, the system can also visualize non-tabular data. By comparison, ShowMe, Voyager, and Draco are domain agnostic tools, that operate under different assumptions compared to GEViTRec. Still, it is instructive to compare between them so as to showcase what GEViTRec is capable of visualizing relative to existing systems. Specifically, this comparison to highlights the advantages (and challenges) of extending recommendation systems multiple tabular and non-tabular data sources that have shared information.

An additional challenge in this comparison is that GEViTRec makes choices about what fields to visualize, optimizing here for generating coordinated visual encoding recommendations across data sources. ShowMe and Draco need the user to provide fields, Voyager does as well but can make field suggestions once the user provides an initial set. GEViTRec can take user input, but does not require it to generate recommendations. It would be prohibitive to show the visualizations ShowMe, Voyager, and Draco could generated with all possible field combinations. We selected a few univariate, bivariate, and multivariate fields for this comparison.

Comparison Data : Ebola Outbreak

We visualize a publicly available Ebola Outbreak dataset, which contains tabular, phylogenetic, and spatial data. All data are available with the GEViTRec code repository. Here is a snapshot of the data:



ebov_metadata

Site_ID	country	location	latitude	longitude	collection_date	year	month
EBOV|20140008|KR653251|SLE|Kenema|2014-08-22	SLE	Kenema	7.790259777	-11.22500302	2014-08-22	2014	8
EBOV|20140024|KR653252|SLE|PortLoko|2014-08-20	SLE	PortLoko	8.656462445	-12.538551	2014-08-20	2014	8
EBOV|20140038|KR653267|SLE|Tonkolili|2014-08-23	SLE	Tonkolili	8.424763876	-12.21726922	2014-08-23	2014	8
EBOV|20140091|KR653239|SLE|WesternRural|2014-08-22	SLE	WesternRural	8.408202778	-13.1496277	2014-08-22	2014	8
EBOV|20140100|KR653241|SLE|Pujehun|2014-08-24	SLE	Pujehun	7.350490502	-11.7520563	2014-08-24	2014	8
EBOV|20140134|KR653227|SLE|Bombali|2014-08-26	SLE	Bombali	8.908954621	-12.05495207	2014-08-26	2014	8
EBOV|20140161|KR653265|SLE|Kailahun|2014-08-27	SLE	Kailahun	8.014160864	-10.7958557	2014-08-27	2014	8
EBOV|20140174|KR653294|SLE|WesternUrban|2014-08-27	SLE	WesternUrban	8.456939201	-13.3212838	2014-08-27	2014	8
EBOV|20140254|KR653296|SLE|Bo|2014-08-29	SLE	Bo	7.988310718	-11.75741265	2014-08-29	2014	8
EBOV|20140300?|KP178538|LBR|?|2014-08-03	LBR	Montserrado	6.305226639	-10.67188092	2014-08-03	2014	8
EBOV|20140395|KR653279|SLE|Kono|2014-09-02	SLE	Kono	8.667233626	-11.03940575	2014-09-02	2014	9
EBOV|20140433|KR653246|SLE|Tonkolili|2014-09-03	SLE	Tonkolili	8.490185653	-12.26573472	2014-09-03	2014	9
EBOV|20140436|KR653287|SLE|Moyamba|2014-09-03	SLE	Moyamba	8.058318704	-12.04242112	2014-09-03	2014	9
EBOV|20140489|KR653235|SLE|Kenema|2014-09-04	SLE	Kenema	7.79034108	-11.21944382	2014-09-04	2014	9
EBOV|20140517|KR653263|SLE|Kailahun|2014-09-05	SLE	Kailahun	8.041925174	-10.83482052	2014-09-05	2014	9
EBOV|20140590|KR653278|SLE|Kenema|2014-09-07	SLE	Kenema	7.850190845	-11.17793679	2014-09-07	2014	9
EBOV|20140729|KR653286|SLE|Bo|2014-09-10	SLE	Bo	7.954879303	-11.7517835	2014-09-10	2014	9
EBOV|20140872|KR653297|SLE|WesternUrban|2014-09-15	SLE	WesternUrban	8.431276717	-13.25748482	2014-09-15	2014	9

Comparison Approach

We visualize only the Tabular Data associated with the Ebola Virus Outbreak Dataset. The Tabular data has 8 fields. We examine only a subset of them:

- Univariate recommended visualizations of all fields (if they are generated)
  - Voyager and Tableau Prep (but not Desktop) generates these automatically
- Bivariate recommended visualizations
  - Country (nominal) and Month (ordinal)
  - Country(nominal) and Collection Date (temporal)
  - Latitude (numeric) and Longitude (numeric)
- Multivariate recommended visualizations
  - Latitude, Longitude, and Month

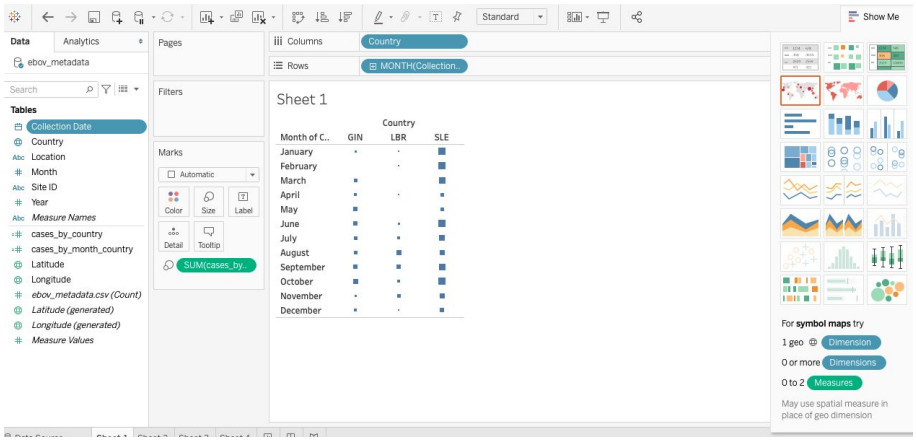
Tableau ShowMe

Default Univariate Summaries

The ShowMe algorithm primarily appears in the Tableau Desktop product, but Desktop does not automatically generate univariate summaries. The Tableau Prep product does produce automatic univariate visualization summaries for fields. For completeness, we include it as part of our comparison.

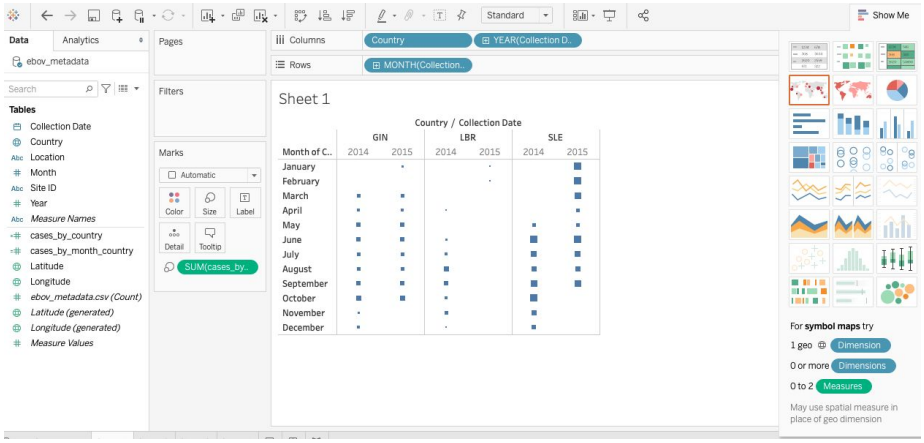
Exploring Bivariate Summaries

Country and Month Fields



Exploring Bivariate Summaries

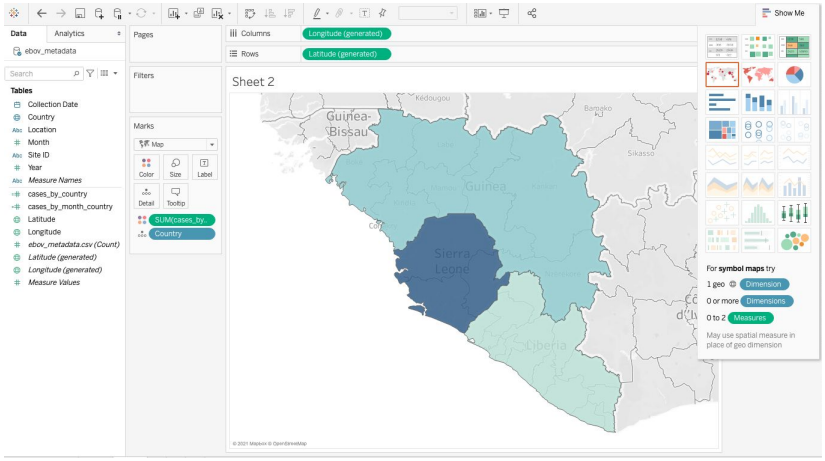
Country and Month Fields



Can easily show collection year as well

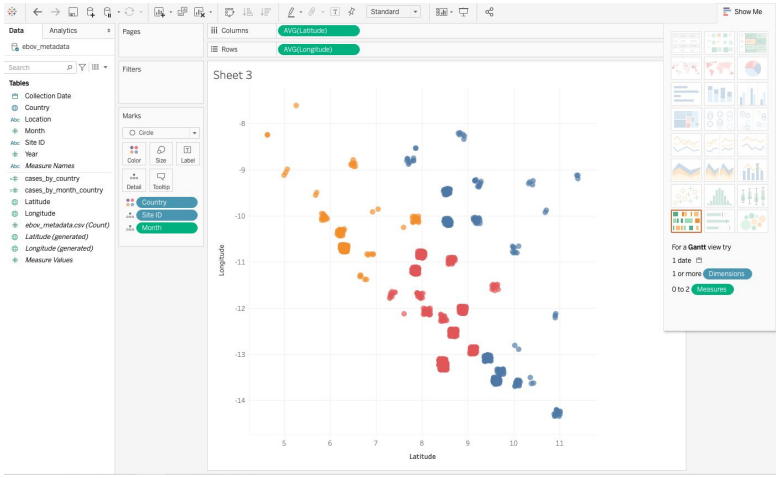
Exploring Bivariate Summaries

Latitude, Longitude



Exploring Bivariate Summaries

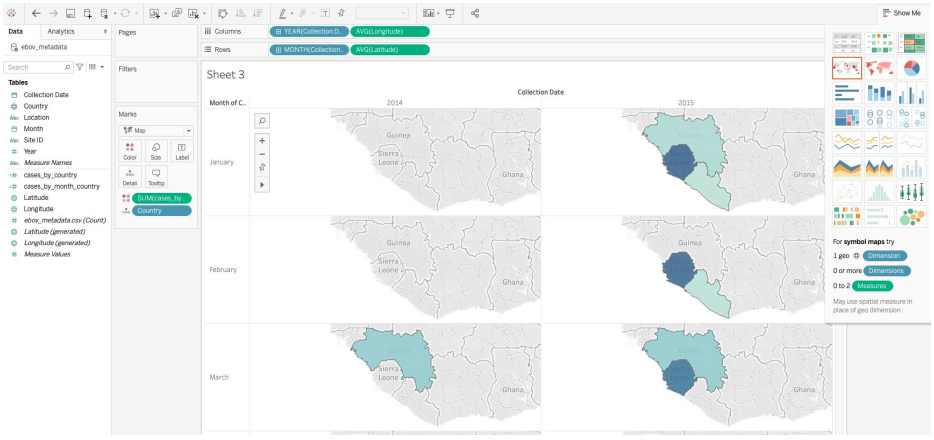
Latitude, Longitude



Without using Tableau's automatically 'generated' coordinates, latitude and longitude can also be treated just as continuous values too.

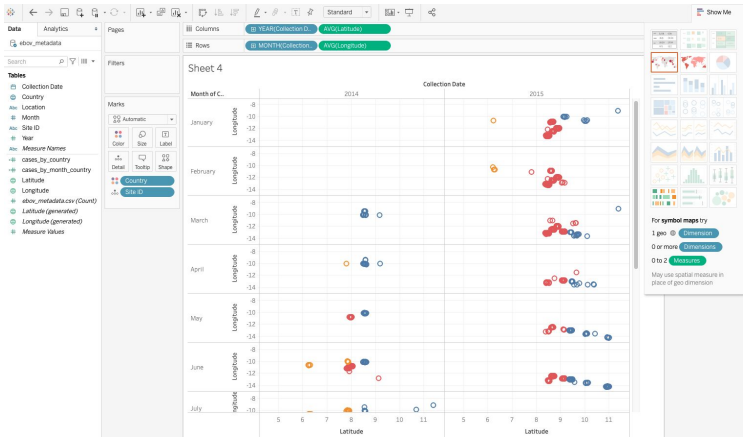
Exploring Multivariate Summaries

Latitude, Longitude, and Month



Exploring Multivariate Summaries

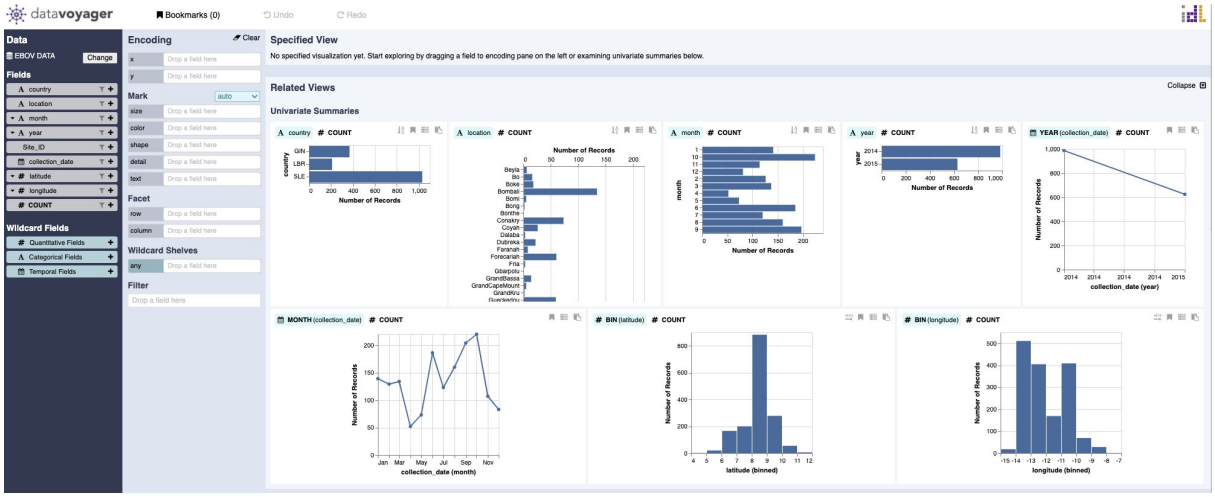
Latitude, Longitude, and Month



Without using Tableau's automatically 'generated' coordinates, latitude and longitude can also be treated just as continuous values too.

Voyager

Default Univariate Summaries



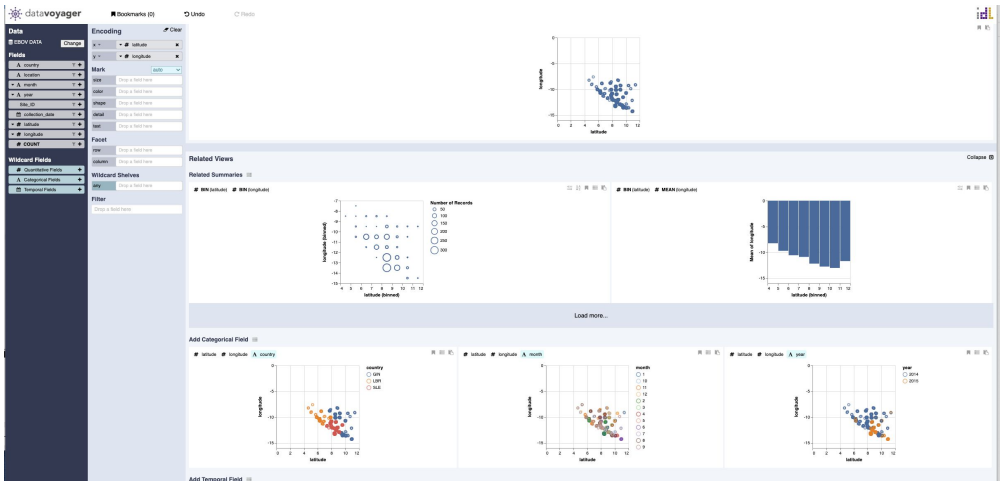
Exploring Bivariate Summaries

Voyager does not make suggestions for bivariate summaries, these must be dragged and dropped different encoding slots. We explore a few different combinations

Country and Month

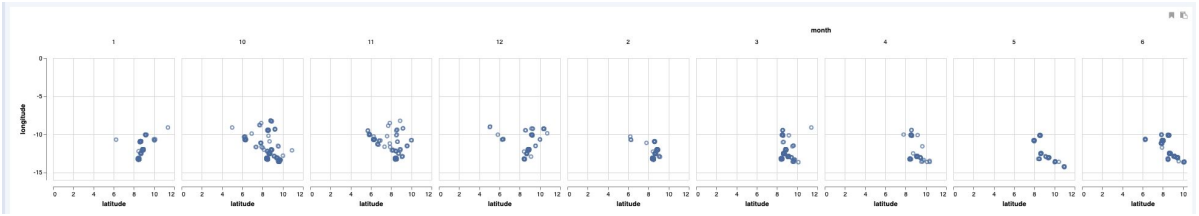
Exploring Bivariate Summaries

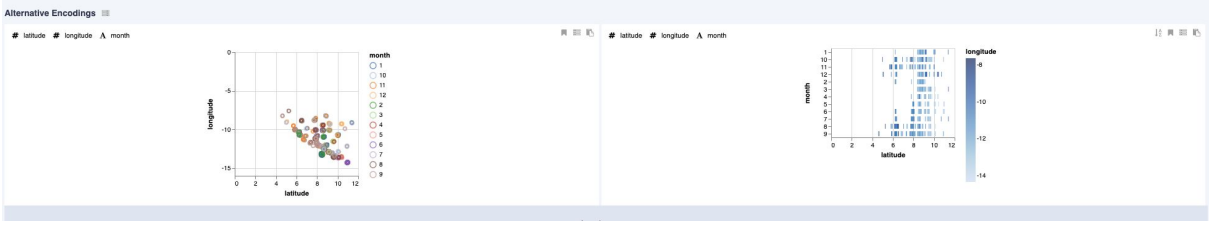
Latitude and Longitude



Exploring Bivariate Summaries

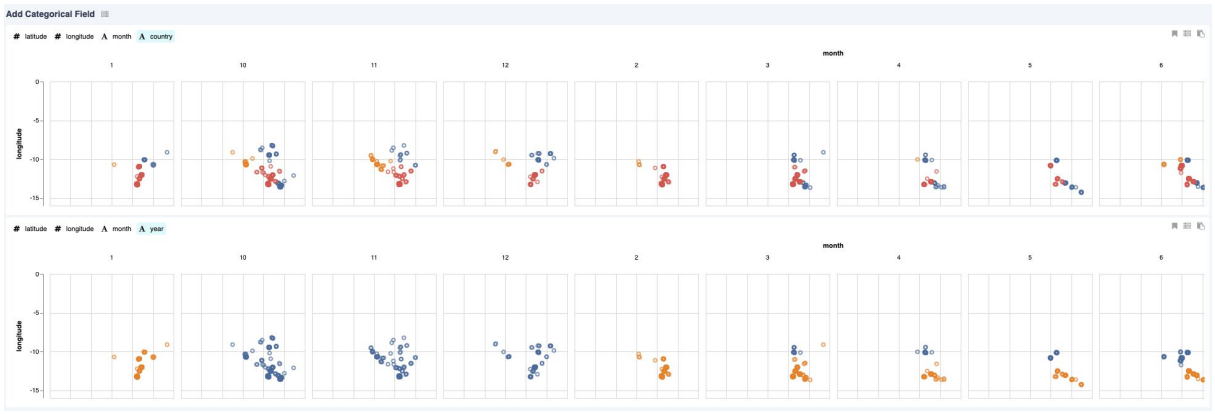
Latitude, Longitude





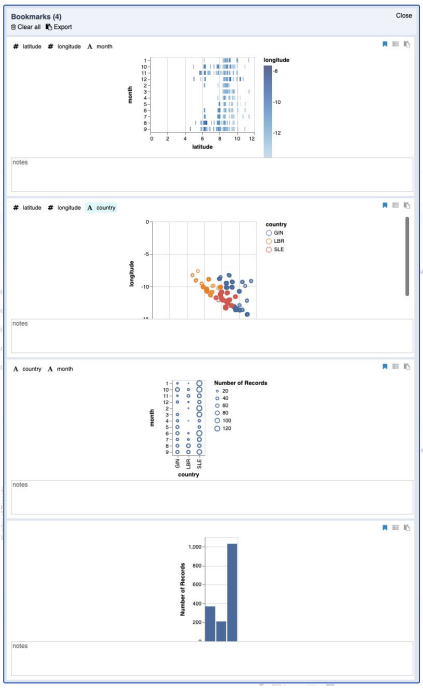
Exploring Multivariate Summaries

Latitude, Longitude, and Month



Exploring Bookmarks

With Voyager, users can also bookmark suggestions that could approximate a ‘supported’, if not totally automated, dashboard creation. Here we have chosen four items that have complementary information and demonstrate how it would be presented by the Voyager system.



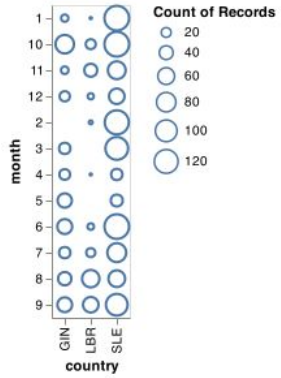
Draco

Default Univariate Summaries

Draco does not produce univariate summaries automatically, however, they can be produced if the user specifies the summaries they wish to generate. Below are some examples.

Exploring Bivariate Summaries

Country and Month Fields

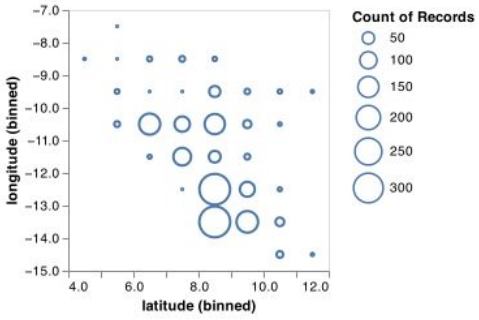


Partial specification

```
% dataset
data("ebov_metadata.csv").

% ====== Data definitions ======

num_rows(1610).

fieldtype(country, string).
fieldtype(month, string).


% ====== Query constraints ======
3 = { encoding(E): encoding(E) }.
```

Exploring Bivariate Summaries

Latitude, Longitude



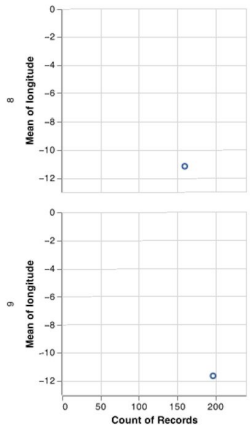
Partial specification

```
% dataset
data("ebov_metadata.csv").

% ====== Data definitions ======

num_rows(1610).

fieldtype(latitude,number).
fieldtype(longitude,number).

% ====== Query constraints ======
3 { encoding(E): encoding(E) } 3.

% y has to be aggregated and
quantitative
:- channel(E,y), not type(E,quantitative).

% x has to be aggregated and
quantitative
:- channel(E,x), not type(E,quantitative).
```

This constraint has to be specified or it doesn't really work

Exploring Multivariate Summaries

Latitude, Longitude, and Month



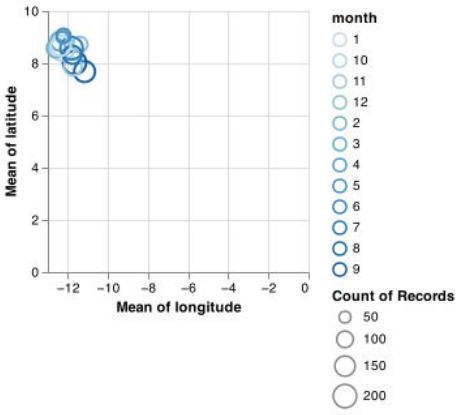
… note this image is cropped. The result is a small multiple display, where each facet (rows) represents one month

Partial specification

```
% dataset for generated VL
data("ebov_metadata.csv").

% ====== Data definitions ======

num_rows(1610).

fieldtype(latitude,number).
fieldtype(longitude,number).
fieldtype(month,string).


% ====== Query constraints ======
3 = { encoding(E): encoding(E) }.


% y has to be aggregated and quantitative
:- channel(E,y), not type(E,quantitative).

% x has to be aggregated and quantitative
:- channel(E,x), not type(E,quantitative).
```

Exploring Multivariate Summaries

Latitude, Longitude, and Month

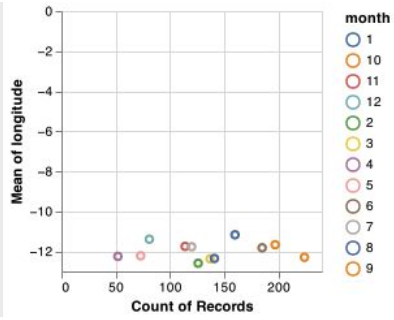


Partial specification - another attempt

```
% dataset for generated VL
data("ebov_metadata.csv").

% ====== Data definitions ======

num_rows(1610).

fieldtype(latitude,number).
fieldtype(longitude,number).
fieldtype(month,string).


% ====== Query constraints ======
2 { type(E,quantitative): encoding(E) } 2.
1 { type(E,nominal): encoding(E) } 1.


% need to use x and y
:- not channel(_,(x;y;color)).

% y has to be aggregated and quantitative
:- channel(E,y), not type(E,quantitative).

% x has to be aggregated and quantitative
:- channel(E,x), not type(E,quantitative).
```

Exploring Multivariate Summaries

Latitude, Longitude, and Month



Partial specification - most constraints

```
% dataset for generated VL
data("ebov_metadata.csv").

% ====== Data definitions ======

num_rows(1610).

fieldtype(latitude,number).
fieldtype(longitude,number).
fieldtype(month,string).


% ====== Query constraints ======
3 { type(E,quantitative): encoding(E) } 3.
1 { type(E,ordinal): encoding(E) } 1.

% need to use x and y
:- not channel(_,(x;y;color)).

% y has to be aggregated and quantitative
:- channel(E,y), not type(E,quantitative).

% x has to be aggregated and quantitative
:- channel(E,x), not type(E,quantitative).

% @constraint Prefer qualitative for bin.
soft(qual_bin,E) :- bin(E,_), not type(E,ordinal).
```

References

[1] Anamaria Crisan, Geoffrey McKee, Tamara Munzner, and Jennifer L. Gardy. Evidence-based design and evaluation of a whole genome sequencing clinical report for the reference microbiology laboratory. *PeerJ*, 6:e4218, January 2018.

[2] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, NY, 2016.

[3] Guangchuang Yu, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk 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.