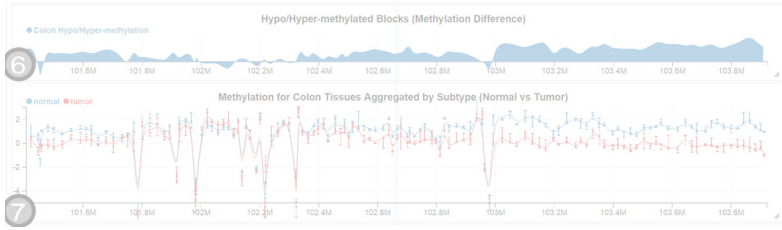


What comes next? Beyond integration of visual and computational interactive data analysis

Héctor Corrada Bravo

@hcorrada

Center for Bioinformatics and Computational Biology
University of Maryland, College Park, USA

CSHL Biological Data Science 2018, 07 November 2018

A longitudinal study

- 30 subjects (8 F, 22M) challenged with enterotoxigenic *E. coli*
- 16S profiling of stool samples one day before, day of and 9 days after
- Diarrehal symptoms recorded
- Ciproflaxin (antibiotic) given after symptoms (or 9 days if no symptoms)

RESEARCH ARTICLE | OPEN ACCESS

Individual-specific changes in the human gut microbiota after challenge with enterotoxigenic *Escherichia coli* and subsequent ciprofloxacin treatment

Mihai Pop, Joseph N. Paulson, Subhra Chakraborty, Irina Astrovskaya, Brianna R. Lindsay, Shan Li, Héctor Corrada Bravo, Clayton Harro, Julian Parkhill, Alan W. Walker, Richard I. Walker, David A. Sack and O. Colin Stine

BMC Genomics 2016 17:440 | https://doi.org/10.1186/s12864-016-2777-0 | © The Author(s). 2016

Received: 8 January 2016 | Accepted: 25 May 2016 | Published: 8 June 2016

Pop et al., 2016 BMC Genomics

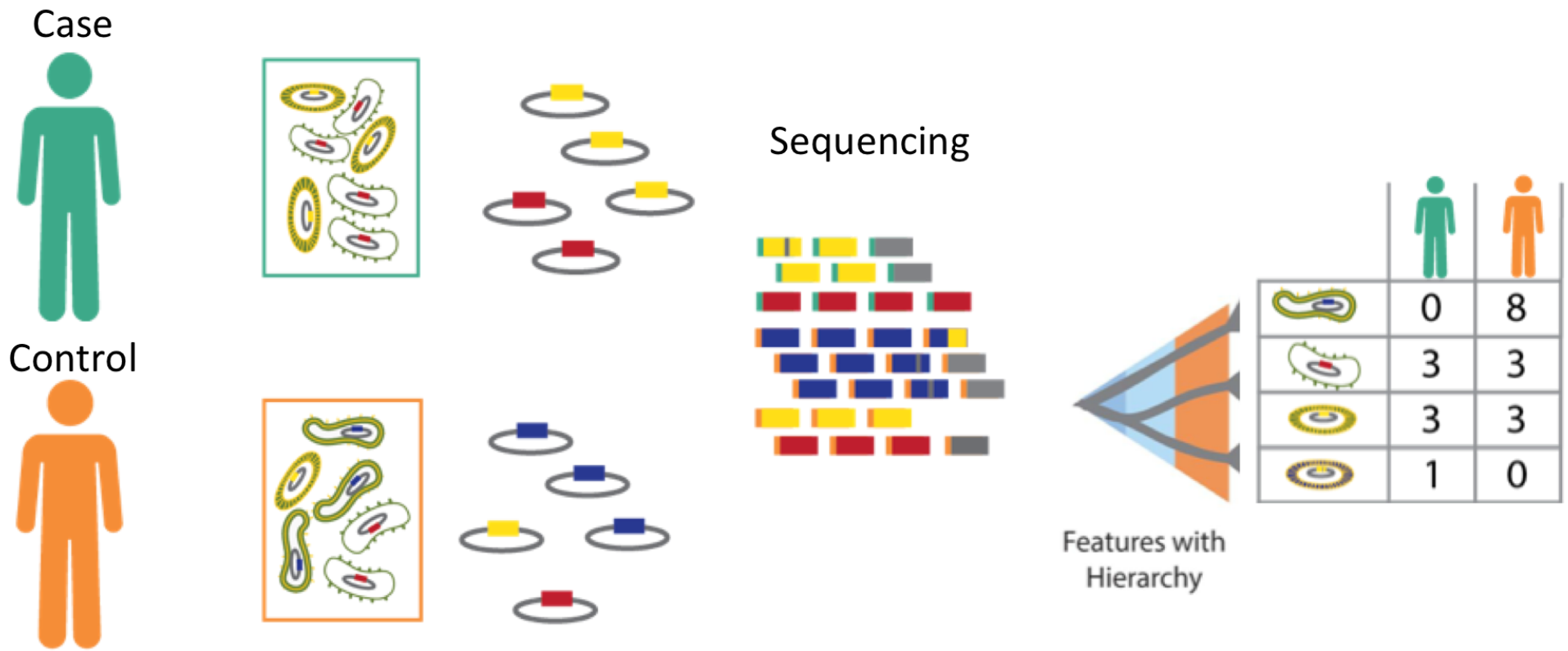
A longitudinal study

Questions

Is there an association between (pre-challenge) microbiome structure and response to challenge

Is there a time structure in the association between specific taxonomic units and presentation of symptoms

Microbiome survey - data generation



This talk

Metaviz

http://metaviz.org

Interactive visualization methods for metagenomic data

Features for longitudinal data in development

This talk

`metagenomeSeq`

http://bioconductor.org/packages/metagenomeSeq

Statistical methods for metagenomic data analysis

Including longitudinal data

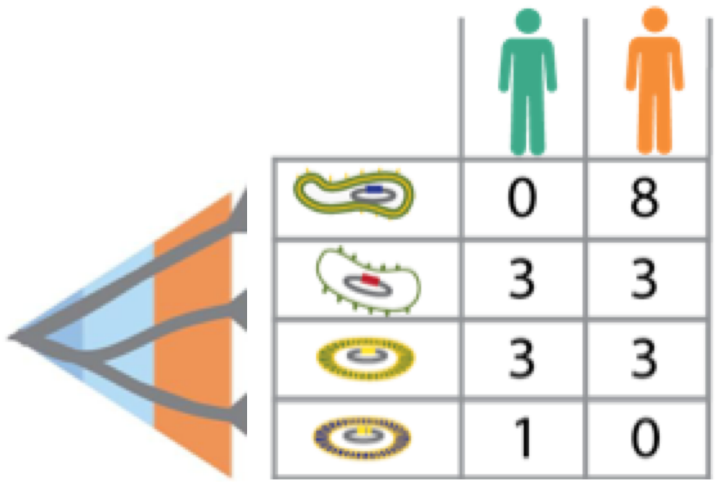
This talk

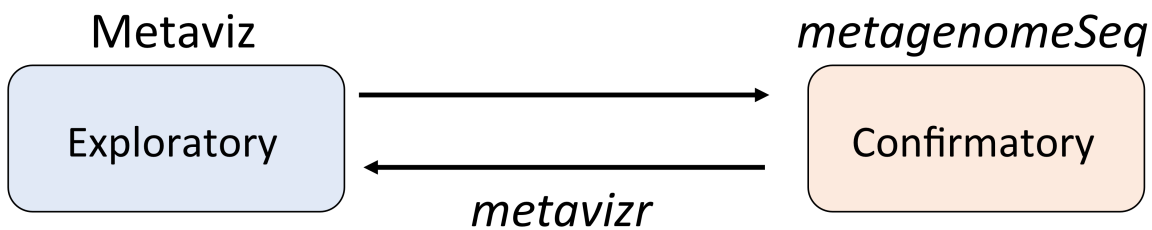
`metavizr`

http://bioconductor.org/packages/metavizr

Integration between R/Bioconductor infrastructure and interactive visualization

This talk


0
8
3
3
3
3
1
0
Features with
Hierarchy


Metaviz
Exploratory
metavizr
metagenomeSeq
Confirmatory

Metaviz

Wagner et al., 2018 Nucleic Acids Research

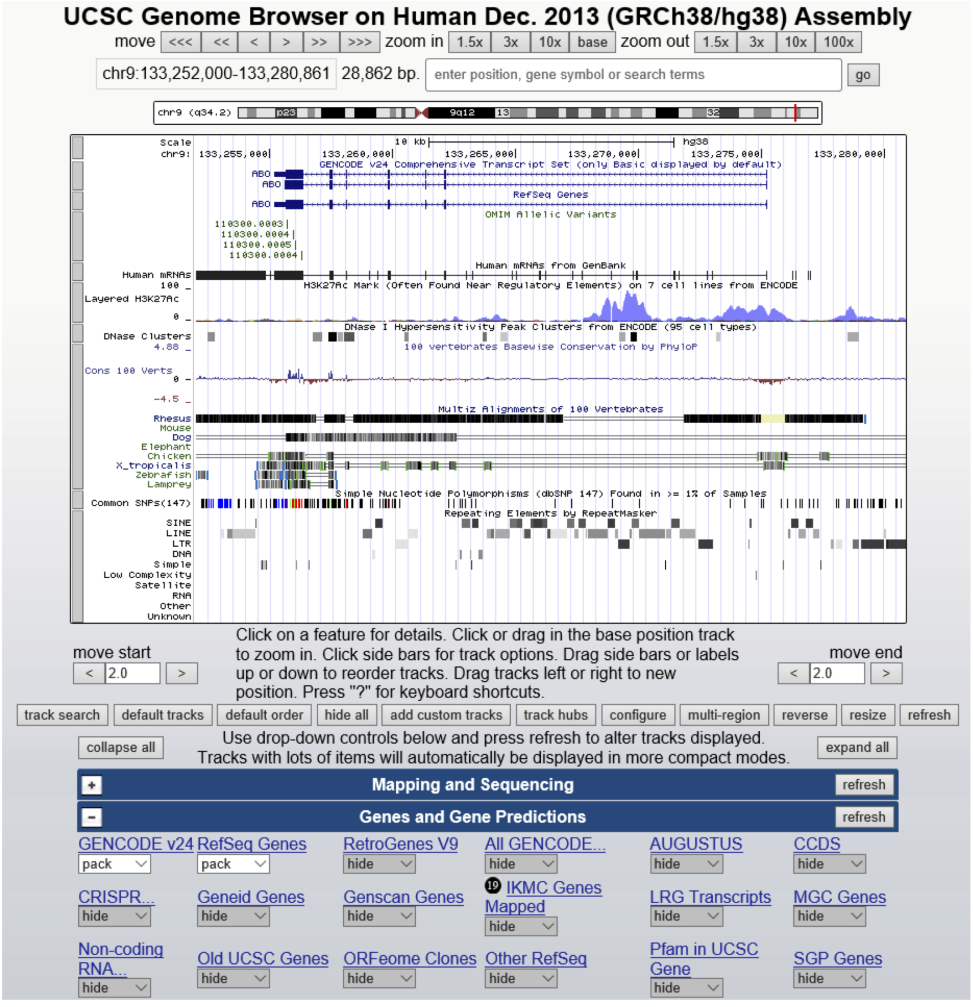
Genome Browsers

Benefits

- Perspective arises naturally to provide quick, intuitive navigation
- Genes and annotations integrated

Track-based

- Sequences and measurements are laid out across screen



Epiviz: Interactive Visualization for Functional Genomics

Separate track-based and non track-based visualization

Data visualizations updated based on user interaction

Integrated with R/Bioconductor

Chelaru et al. 2014, Nat. Methods

Microbiome data visualization challenges

Features with a hierarchy

- Need to view multiple levels of hierarchy made of thousands of nodes

Need to support population surveys of many samples

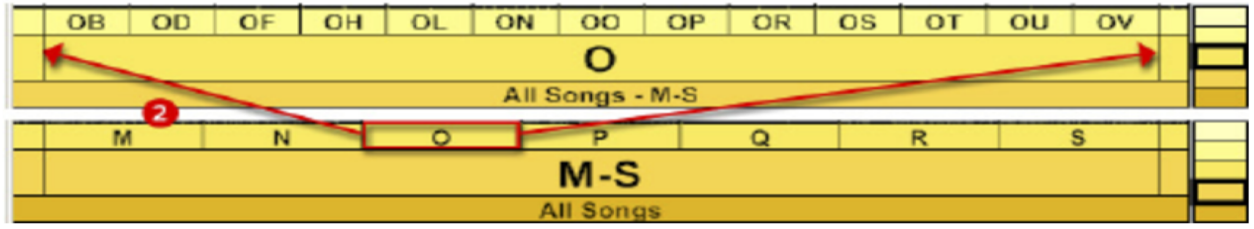
Support quantitative measurements with heatmaps, scatterplots, and boxplots

Hierarchical navigation

Information Visualization technique for hierarchical data

Examine data at different levels and perform feature selection

FacetZoom



Zoom to lower levels

Subset of Levels

Root of Hierarchy

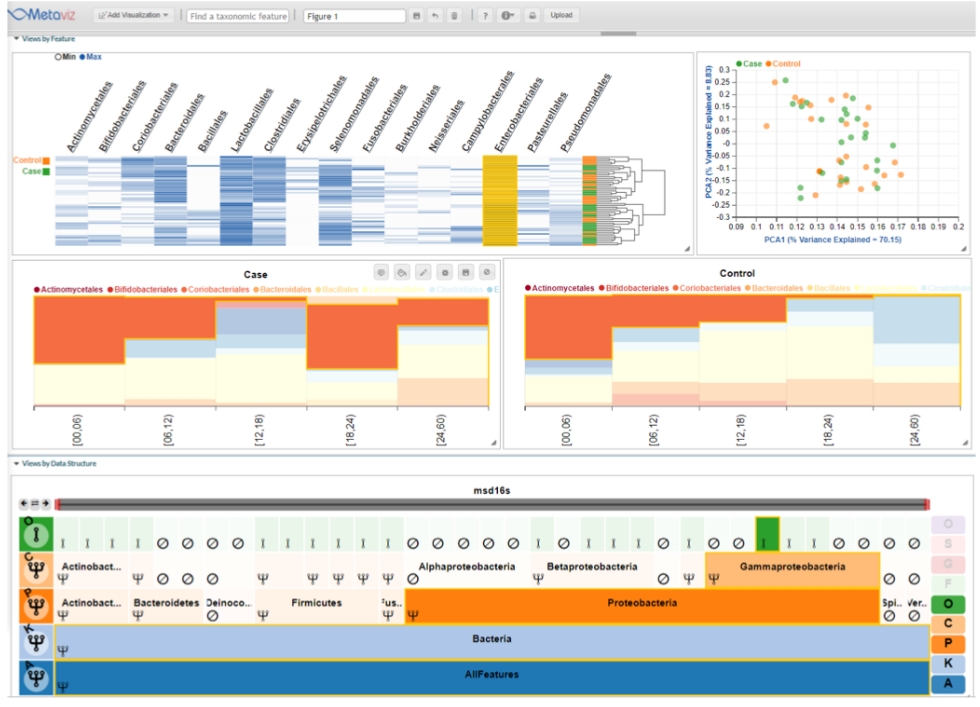
Daschelt et al. 2008. CHI

Metaviz

Integrated, interactive taxonomic feature selection

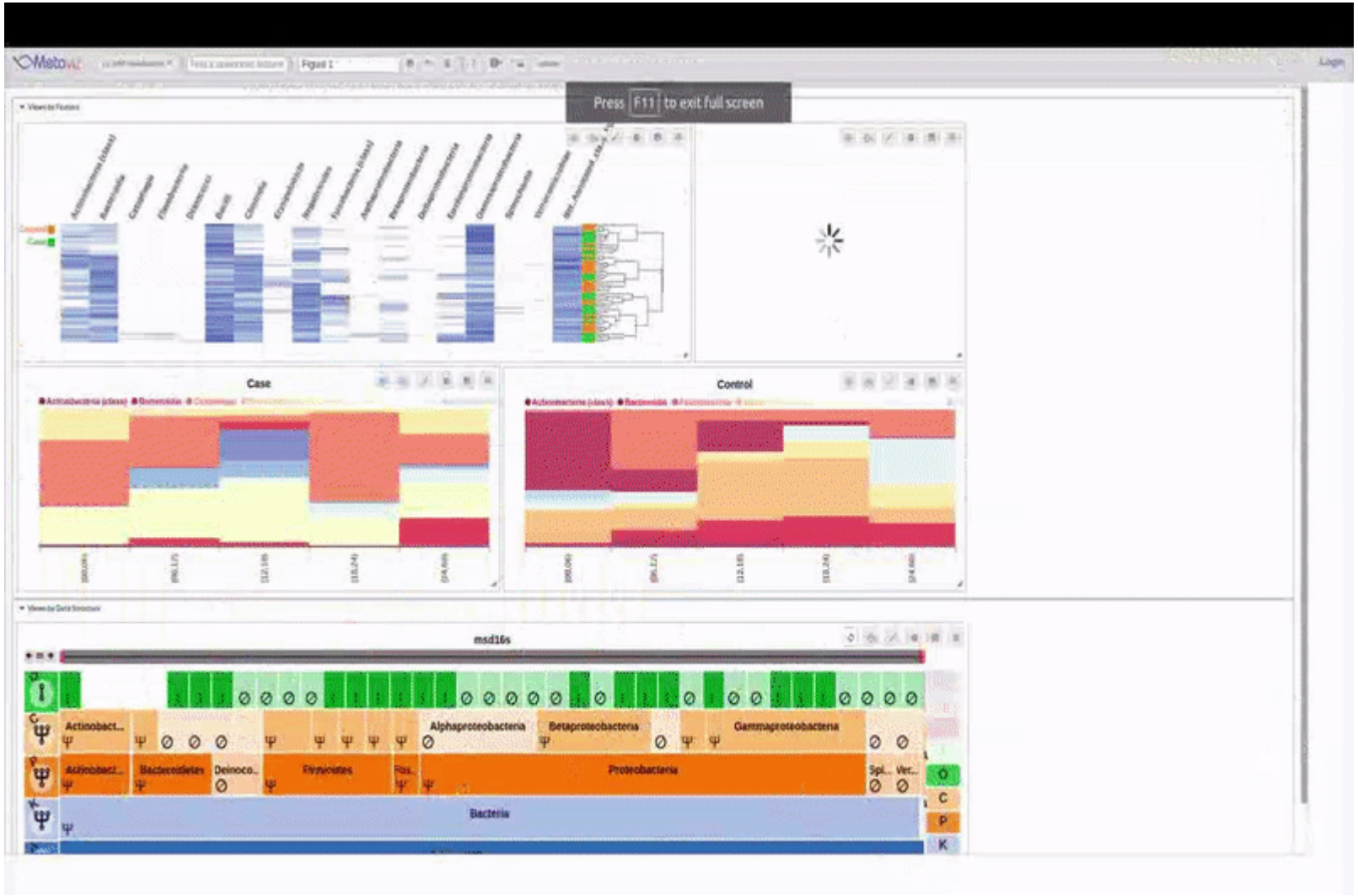
Statistically guided visual analysis

Intuitive navigation for mechanism for hierarchical metagenomic data



Wagner et al., 2018. Nucleic Acids Research

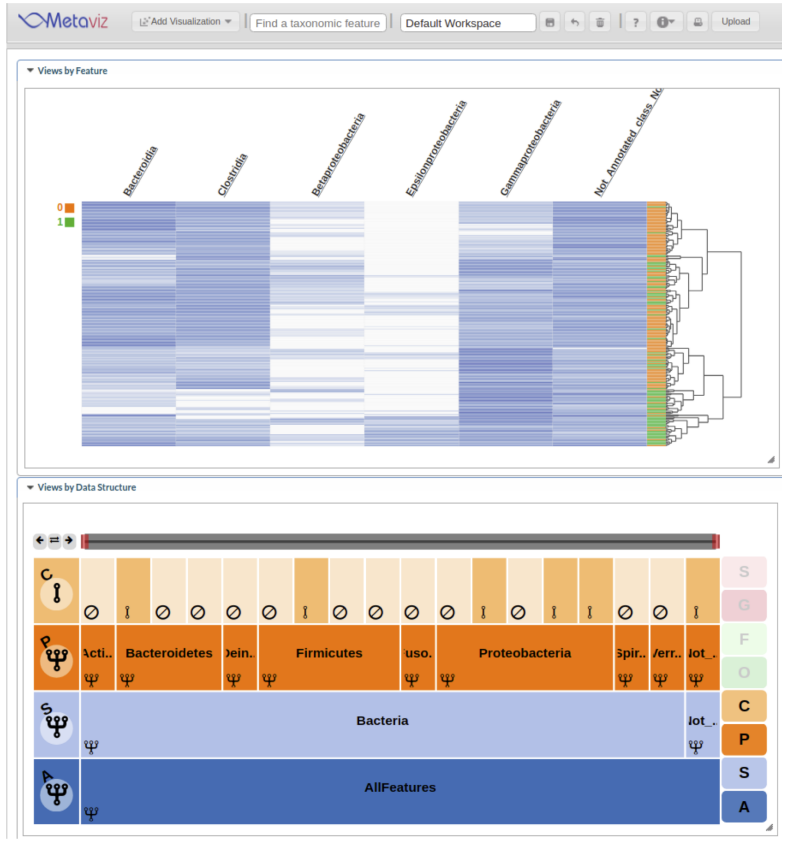
Metaviz


Press F11 to exit full screen
Case
Control
msd16s
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Proteobacteria
Bacteria

Integration with Bioconductor

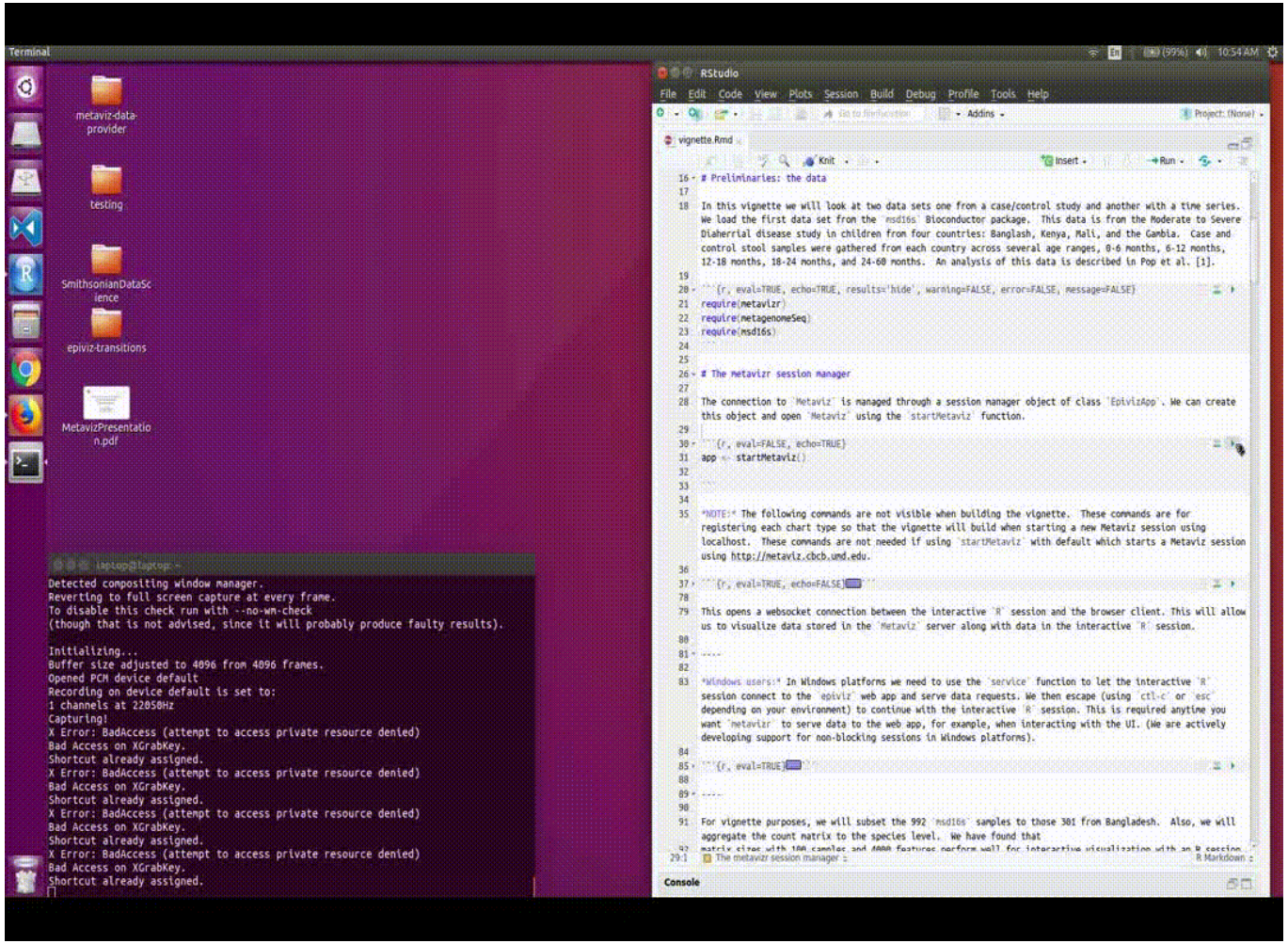
Compute differential abundance with `metagenomeSeq`

Explore results through interactive visualization

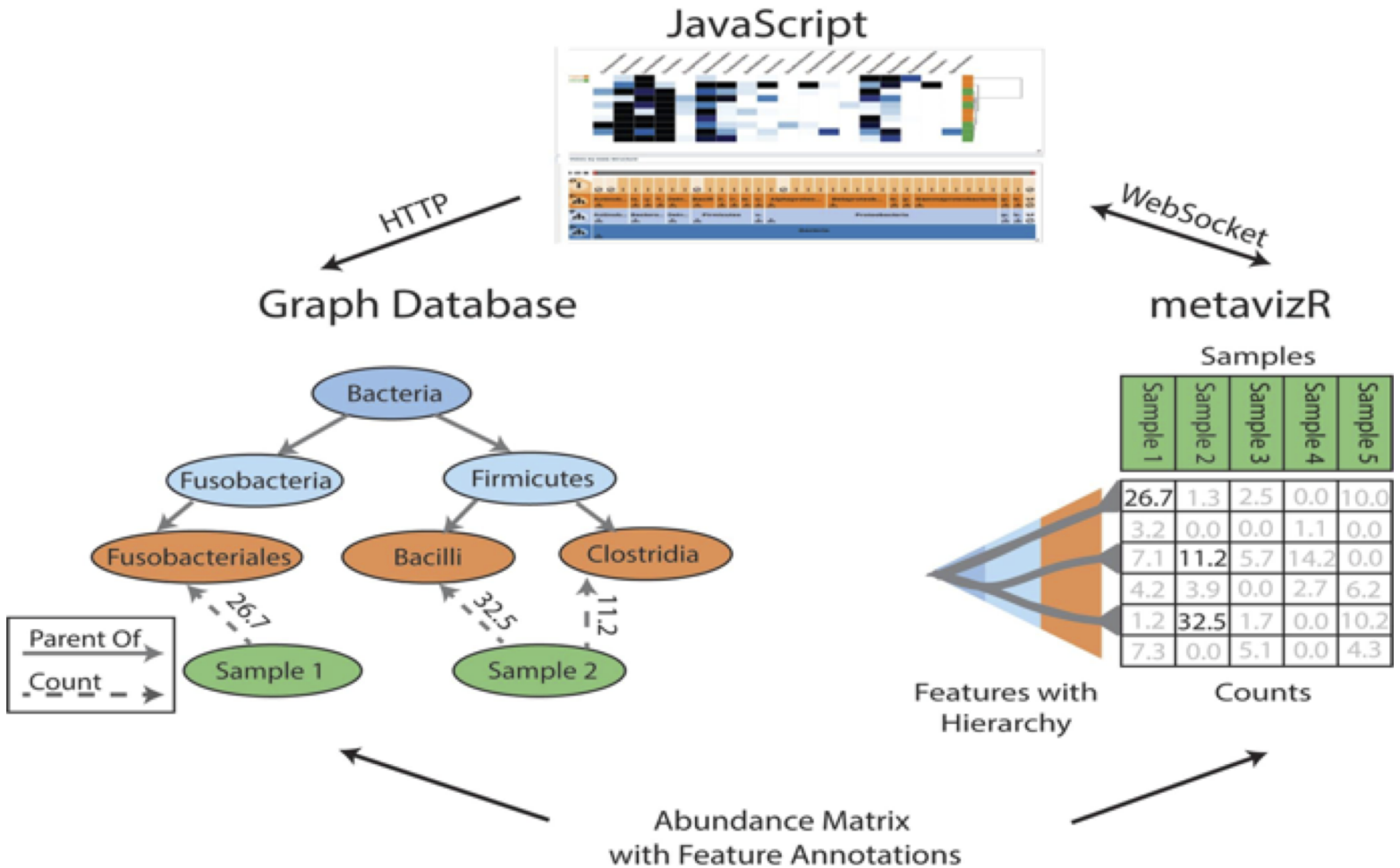


http://bioconductor.org/packages/metavizr

Integration with Bioconductor



Architecture



Deployment options

University of Maryland Metagenome Browser (14k metagenomes)

http://metaviz.cbcb.umd.edu

`metavizr`

http://bioconductor.org/packages/metavizr.html

Local install

https://github.com/epiviz/Metaviz

https://github.com/epiviz/metaviz-data-provider

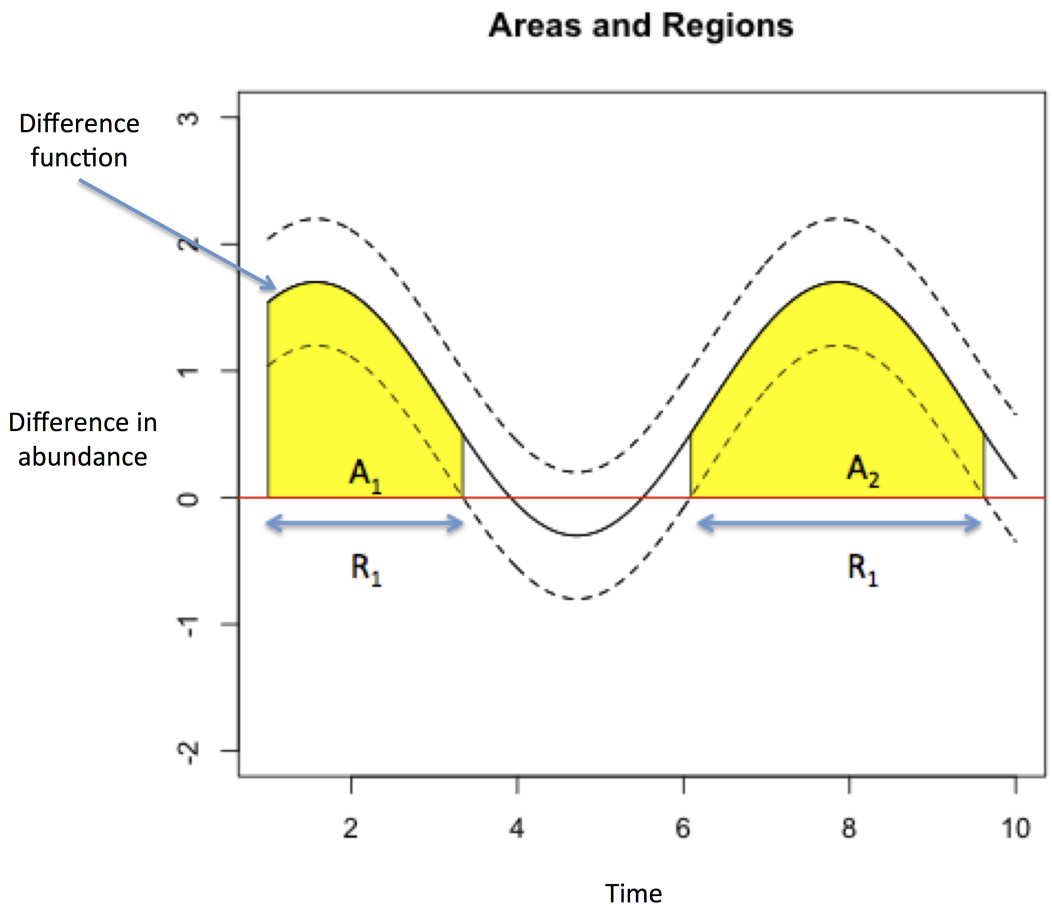
metagenomeSeq

Paulson et al., 2017. biorxiv

Paulson et al., 2013. Nat. Methods

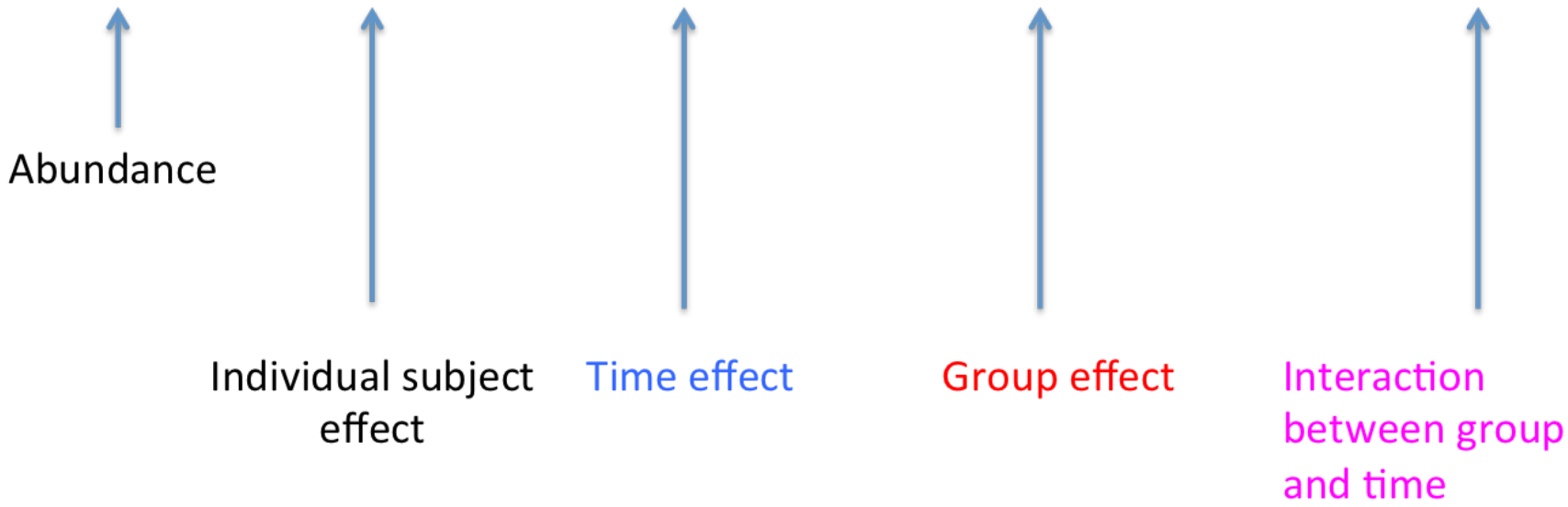
Statistical modeling

Detect intervals of differential abundance between two groups of interest.



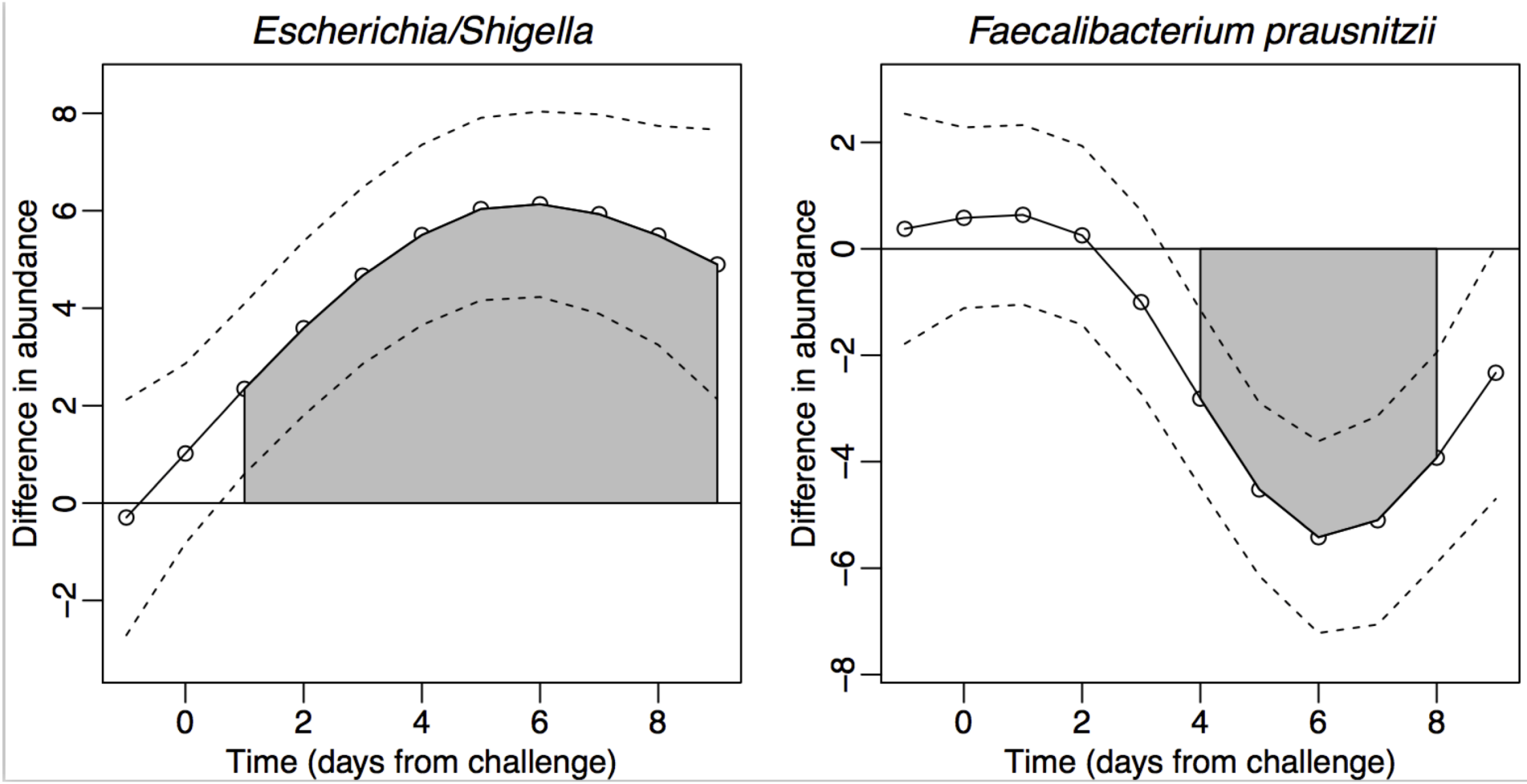
Statistical modeling

$$f_i(t, x_k) = \beta^T x_k + f_1(t) + f_2(I\{k \in i\}) + f_{12}(t, I\{k \in i\})$$



Fit using smoothing-spline ANOVA (`gss` package)

Statistical modeling

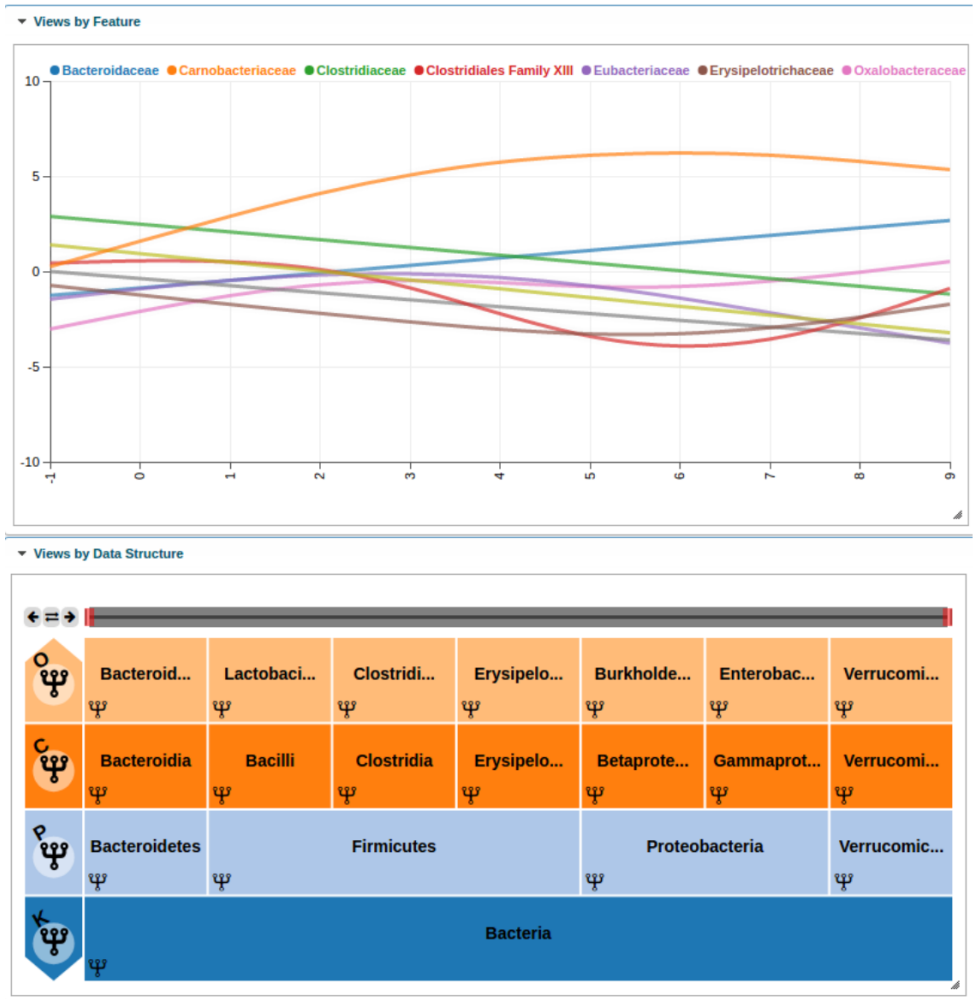


metavizr

http://bioconductor.org/packages/metavizr

Interactivity with spline model

```
library(metavizr)

library(etec16s)

data(etec16s)


app <- startMetaviz()

app$plot(etec16, type="TimeSeries",

         formula=abundance~id+AntiG

           time*AnyDayDiarrhea

)
```



Interactivity with spline model

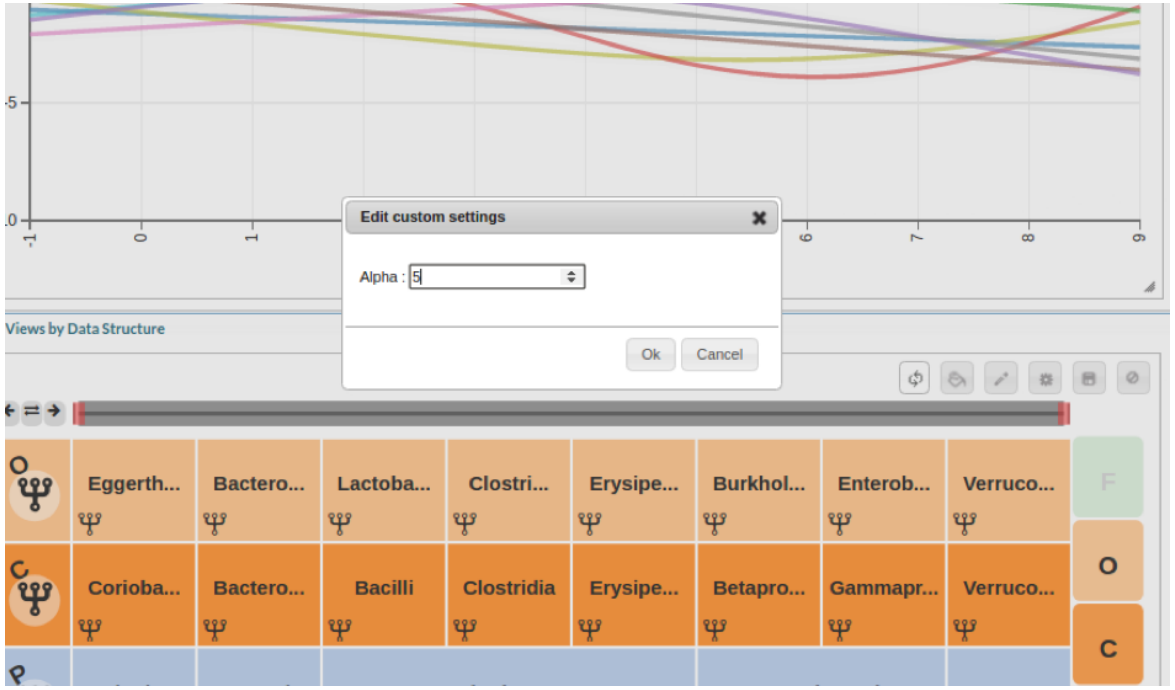
Inference depends on model parameter (smoothness of function)

Use interactivity to explore sensitivity

Interactivity with spline model

Inference depends on model parameter (smoothness of function)

Use interactivity to explore sensitivity

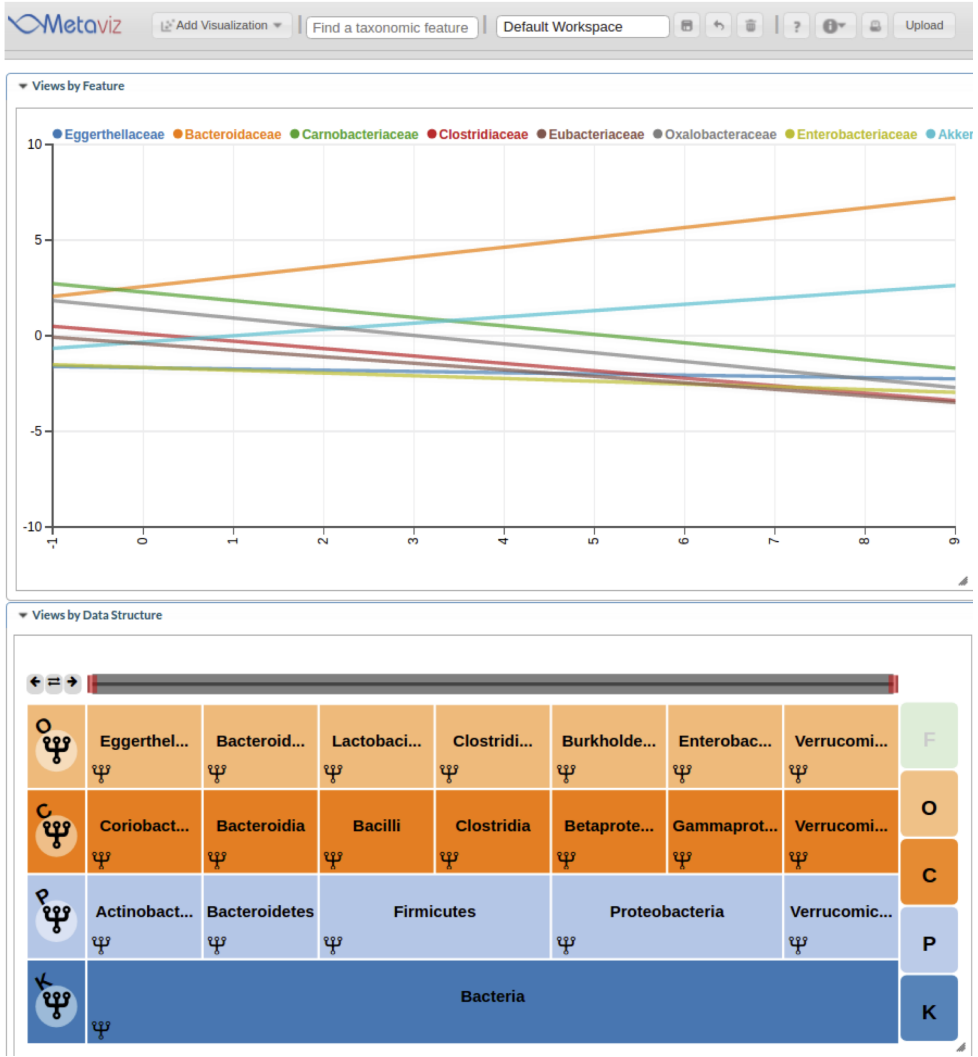


Interactivity with spline model

Inference depends on model parameter (smoothness of function)

Use interactivity to explore sensitivity

High smoothness penalty

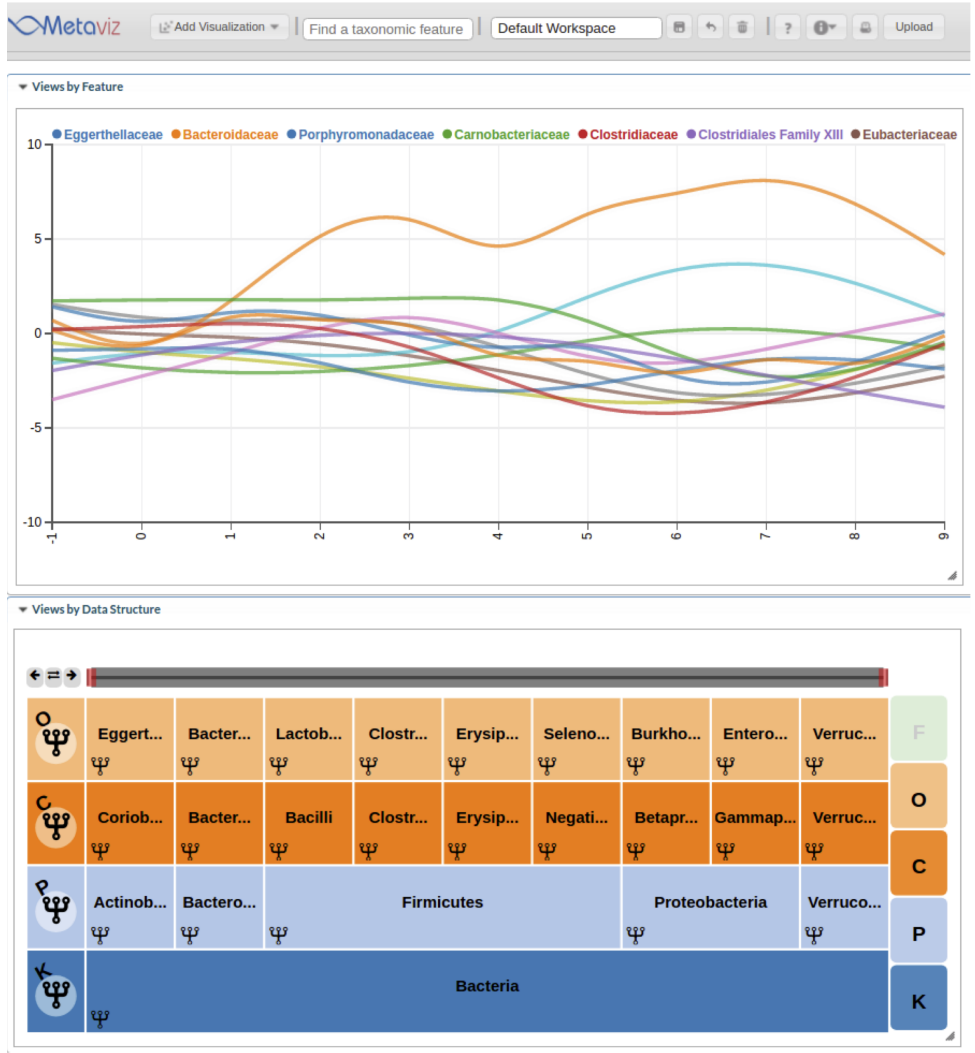


Interactivity with spline model

Inference depends on model parameter (smoothness of function)

Use interactivity to explore sensitivity

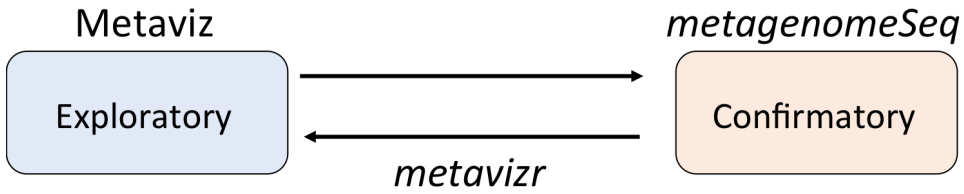
Low smoothness penalty



Statistically guided visualization

Let's revisit the Data Analysis modes

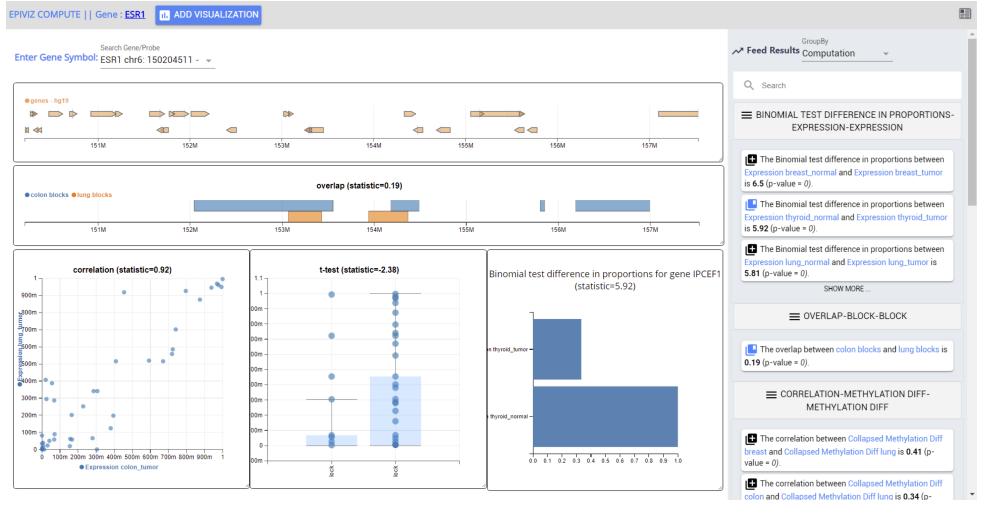
How can we make the connection between two modes of analysis tighter?



Statistically guided visualization

One idea: statistically guided visual exploration via *proactive computation*

Back to Epiviz: epigenetic regulation of gene expression across multiple tumor types (Timp et al. 2016)

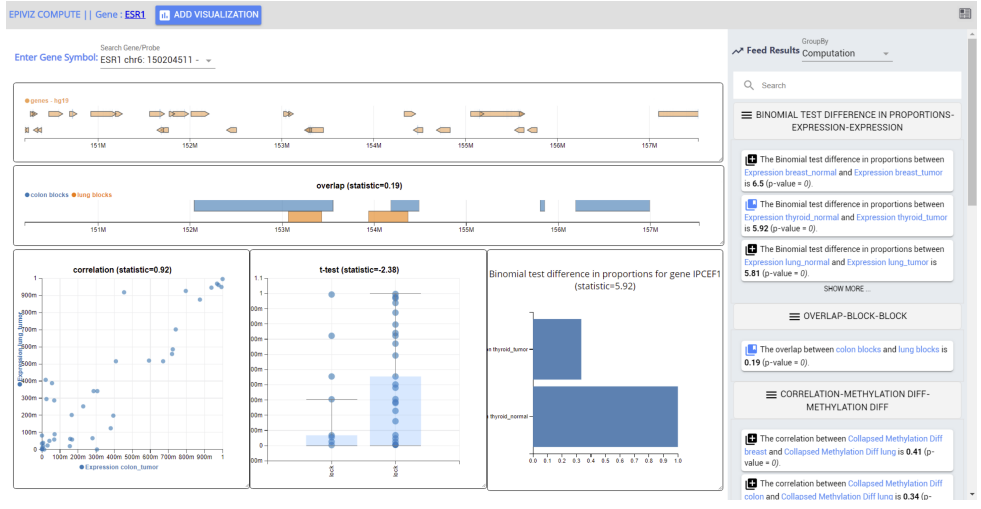


Statistically guided visualization

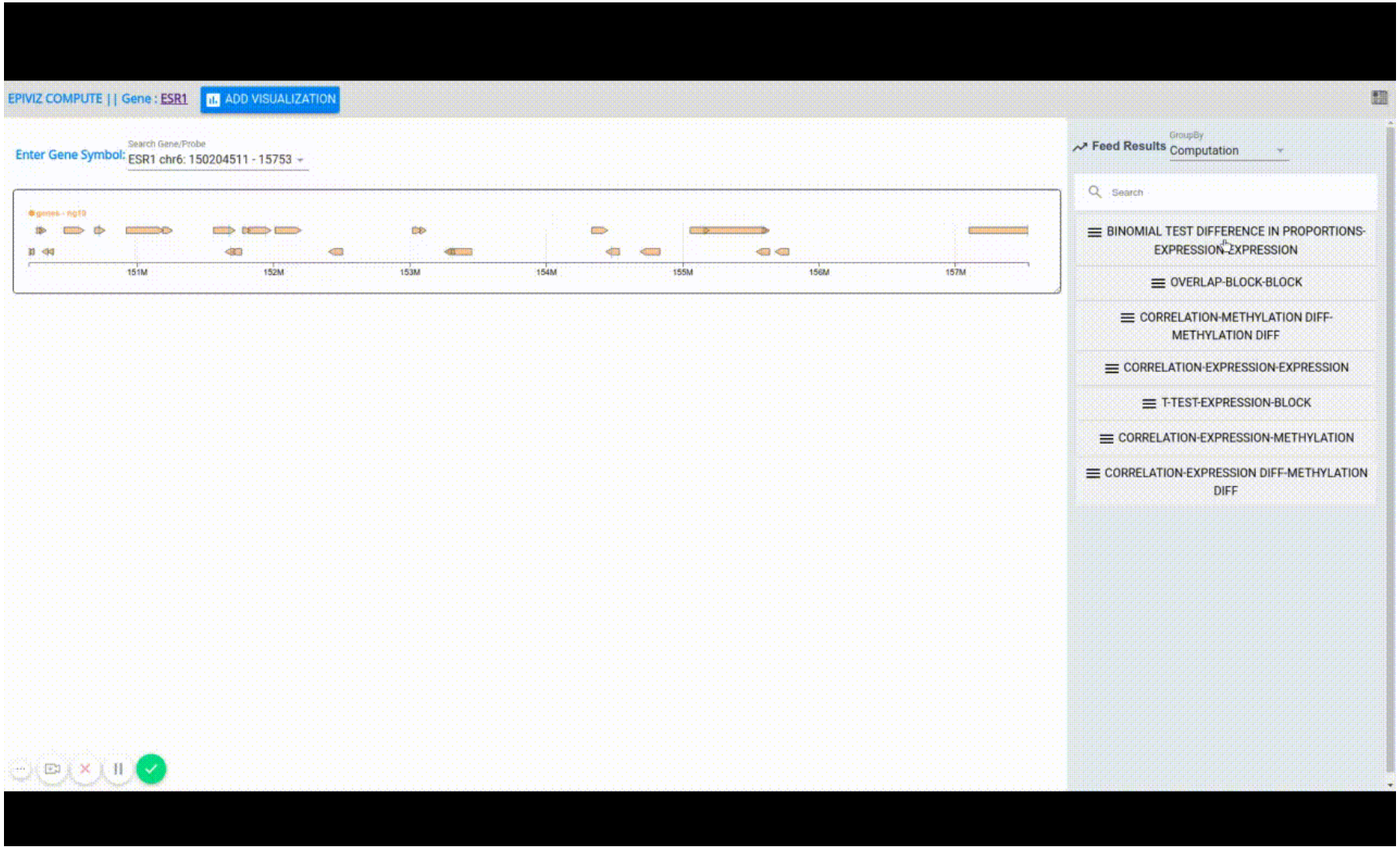
One idea: statistically guided visual exploration via *proactive computation*

Analysis runs in background (proactive computation)

"Interesting results" presented to user, guides exploration of data underlying result.



Statistically guided visualization

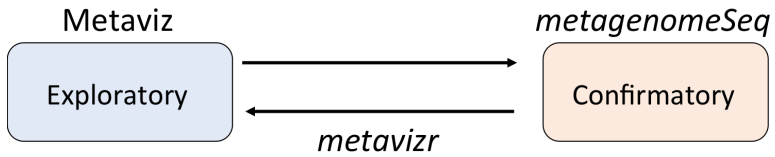


Interactive Data Analysis

In general, how do we support interactive data analysis?

Workflow construction: keep pieces small, we like to do that for *reusability*, now let's do it for *data reflection*.

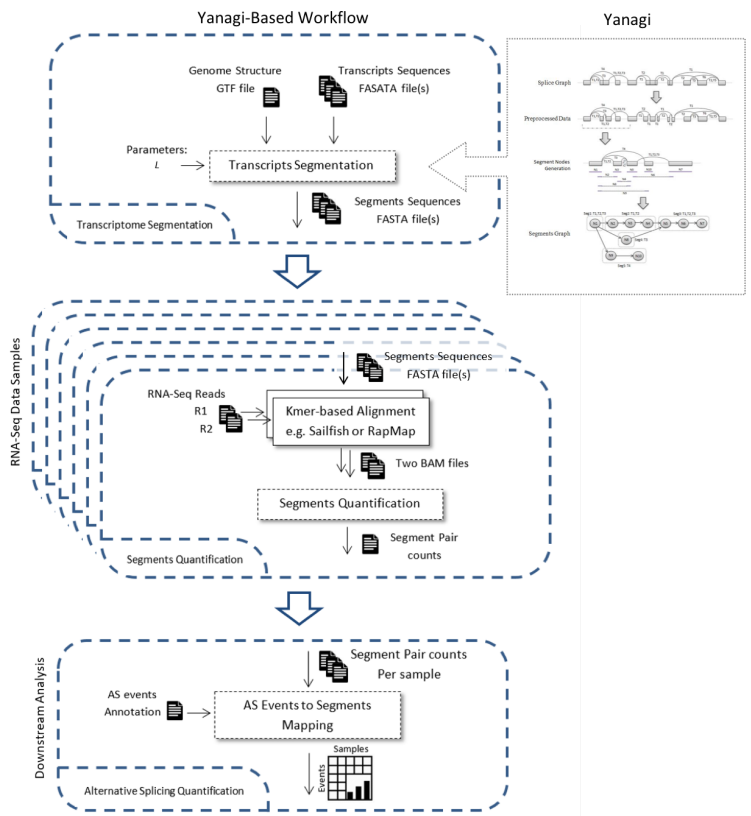
What is the "sufficient statistic" my tool produces that a user can interact with?



Interactive Data Analysis

One example: segment-based analysis of RNA-seq data

Transcript quantification as sufficient statistic?

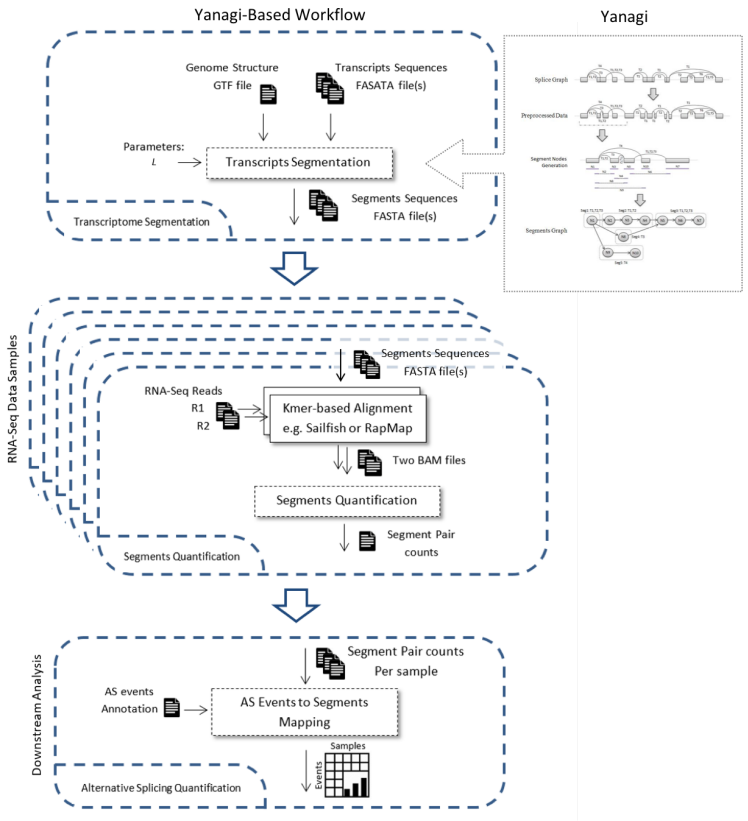


Interactive Data Analysis

One example: segment-based analysis of RNA-seq data

No. There is (pseudo)-counting, and there is quantification.

Can we produce a "sufficient statistic" we can interact with *before* quantification.

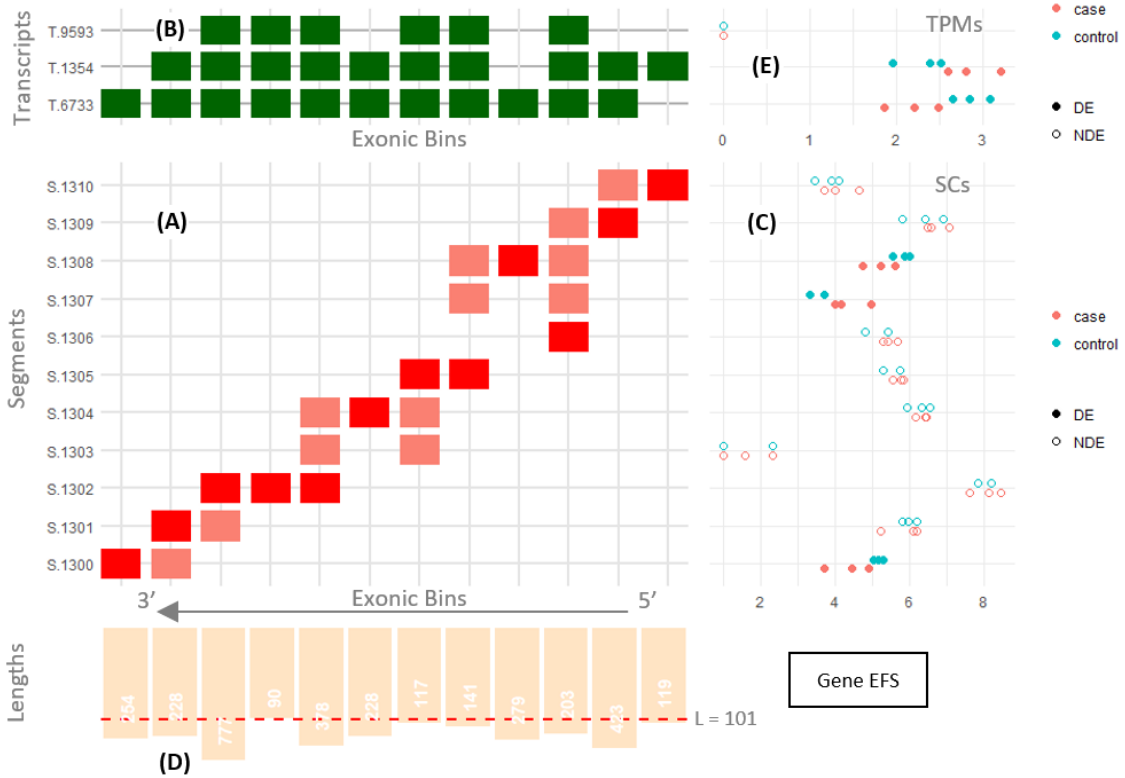


Interactive Data Analysis

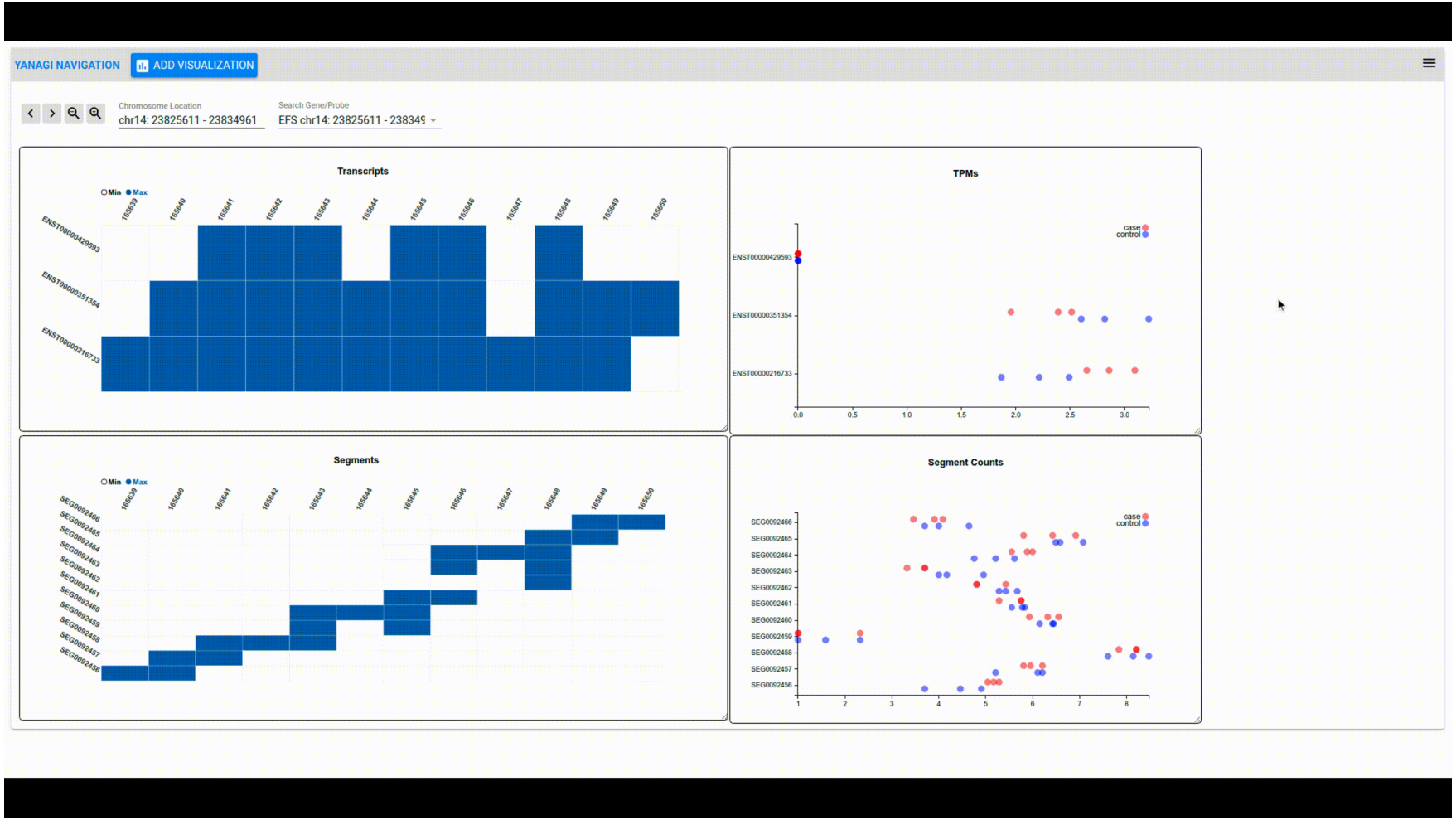
One example: segment-based analysis of RNA-seq data

Segment-counts as the "sufficient statistic"

Gunady et al. Yanagi (2018) biorxiv



Interactive Data Analysis



Summary

Exploration of hierarchical features (FacetZoom)

Summary

Exploration of hierarchical features (FacetZoom)

Design for longitudinal experimental structure (SparklineMatrix)

Summary

Exploration of hierarchical features (FacetZoom)

Design for longitudinal experimental structure (SparklineMatrix)

Interact with data and statistical model via visualization (`metavizr`)

Summary

Exploration of hierarchical features (FacetZoom)

Design for longitudinal experimental structure (SparklineMatrix)

Interact with data and statistical model via visualization (`metavizr`)

Design your tools with data reflection, build IDA tools around that

Acknowledgements

CBCB/UMD College Park

Justin Wagner, Joseph Paulson, Jayaram Kancherla, Florin Chelaru
Mihai Pop, Niklas Elmqvist, Zhe Cui,
Mohamed Gunady, Stephen Mount

UMD Baltimore

Brianna Lindsey, O. Colin Stine, Owen White, Anup Mahurkar

Funding

NIH/NIGMS, Genentech

More info

http://epiviz.org
http://www.hcbravo.org