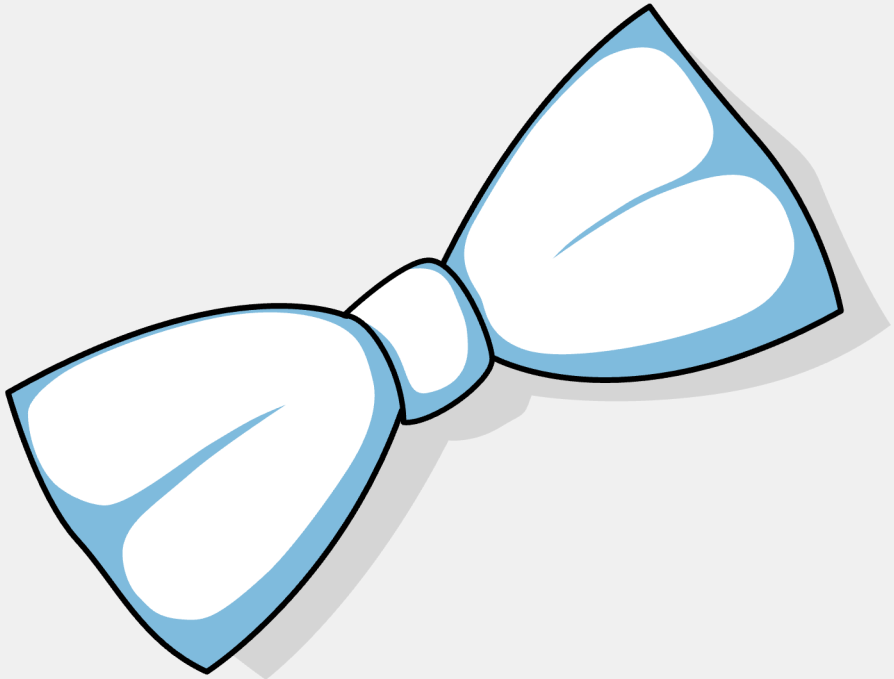


GenomeSpy Workshop

At Oslo Bioinformatics Workshop Week 2024

11.12.2024

About Me

- Kari Lavikka
- PhD student at Hautaniemi Lab,
  Medical Faculty, University of Helsinki
- MSc in Data Science, BSc in Biology
- Author of GenomeSpy

- Email: kari.lavikka@helsinki.fi
- Website: https://karilavikka.fi/

Hautaniemi Lab

OXFORD

(GIGA)n SCIENCE

GigaScience, 2024, **13**, 1–15

DOI: 10.1093/gigascience/giae040

Technical Note

Deciphering cancer genomes with GenomeSpy: a grammar-based visualization toolkit

Kari Lavikka [1,*], Jaana Oikkonen [1], Yilin Li [1], Taru Muranen [1], Giulia Micoli [1], Giovanni Marchi [1], Alexandra Lahtinen [1], Kaisa Huhtinen [1,2], Rainer Lehtonen [3], Sakari Hietanen [4], Johanna Hynninen [4], Anni Virtanen [5], and Sampsa Hautaniemi [1,*]

[1]Research Program in Systems Oncology, Research Programs Unit, Faculty of Medicine, University of Helsinki, 00014 Helsinki, Finland
[2]Cancer Research Unit, Institute of Biomedicine and FICAN West Cancer Centre, University of Turku, 20521 Turku, Finland
[3]Applied Tumor Genomics Research Program, Research Programs Unit, Faculty of Medicine, University of Helsinki, 00014 Helsinki, Finland
[4]Department of Obstetrics and Gynecology, University of Turku and Turku University Hospital, 20521 Turku, Finland
[5]Department of Pathology, University of Helsinki and HUS Diagnostic Center, Helsinki University Hospital, 00260 Helsinki, Finland
*Correspondence address. Kari Lavikka, Research Program in Systems Oncology, Research Programs Unit, Faculty of Medicine, University of Helsinki, 00014 Helsinki, Finland, E-mail: kari.lavikka@helsinki.fi; Sampsa Hautaniemi, Research Program in Systems Oncology, Research Programs Unit, Faculty of Medicine, University of Helsinki, 00014 Helsinki, Finland, E-mail: sampsa.hautaniemi@helsinki.fi

Abstract

Background: Visualization is an indispensable facet of genomic data analysis. Despite the abundance of specialized visualization tools, there remains a distinct need for tailored solutions. However, their implementation typically requires extensive programming expertise from bioinformaticians and software developers, especially when building interactive applications. Toolkits based on visualization grammars offer a more accessible, declarative way to author new visualizations. Yet, current grammar-based solutions fall short in adequately supporting the interactive analysis of large datasets with extensive sample collections, a pivotal task often encountered in cancer research.

Findings: We present GenomeSpy, a grammar-based toolkit for authoring tailored, interactive visualizations for genomic data analysis. By using combinatorial building blocks and a declarative language, users can implement new visualization designs easily and embed

Program

- Background: DECIDER project
- Background: Why yet another tool
- SegmentModel Spy demo
- DECIDER visualization demo and exercise
- Visualization Grammars
- Code-Along: Let's make a ClinVar visualization
- Using in Observable notebooks
- Local development, remote deployment
- Questions, discussion, etc.

Survey

Background

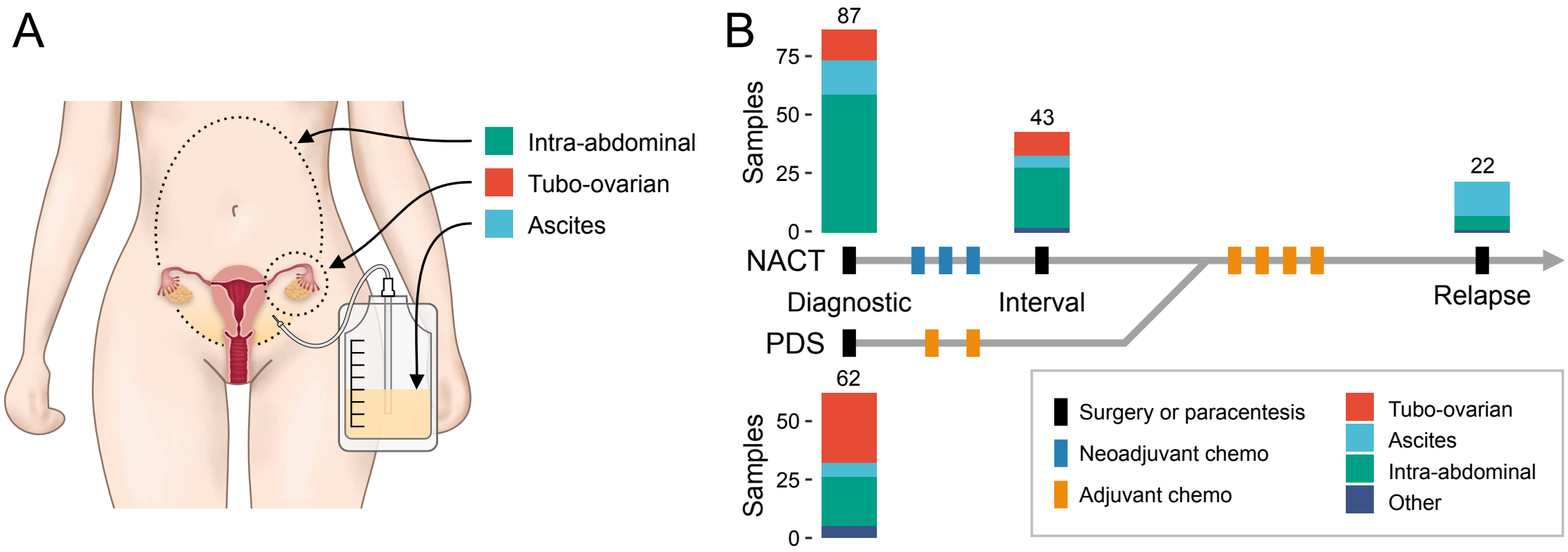
Why was GenomeSpy developed?

DECIDER

Improving clinical decisions in cancer

- Ovarian High-Grade Serous Carcinoma (HGSC)
- Diagnostic tools and treatments
- Overcoming Drug Resistance
- Funded by Horizon 2020

DECIDER HGSC Cohort



A. Lahtinen et al., "Evolutionary states and trajectories characterized by distinct pathways stratify patients with ovarian high grade serous carcinoma," *Cancer Cell*, May 2023, doi: 10.1016/j.ccell.2023.04.017.

Geneticists and others work with...

- ~1500 WGS samples
- Copy numbers
- SNVs / Indels
- Structural Variants
- RRBS Methylation
- Expression data

Geneticists and others need to...

- Explore the whole cohort for patterns
- Stratify the cohort in various ways
- Study individual patients
- Rapidly check hypotheses
- Etc...

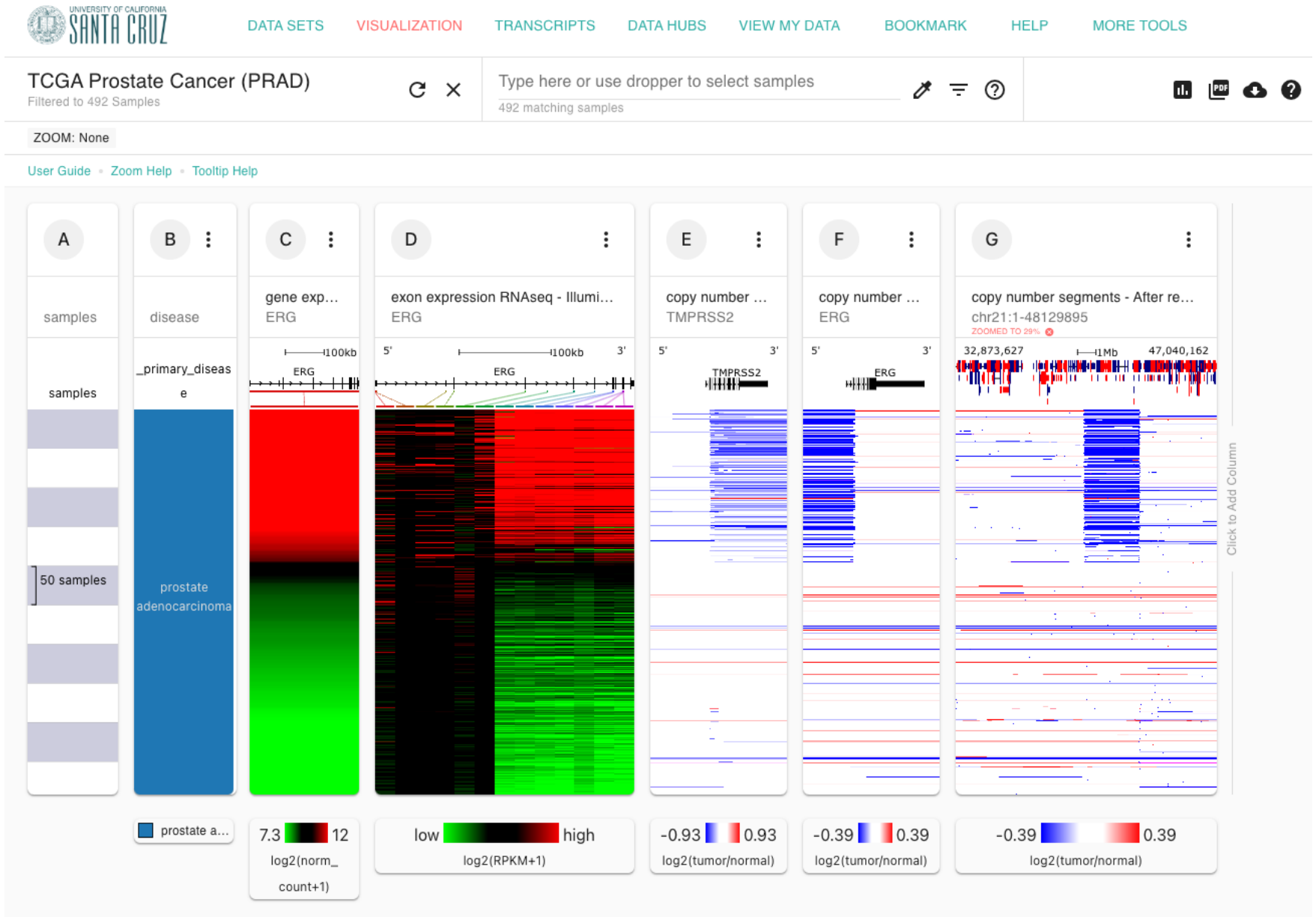
The State of the Art

awesome-genome-visualization

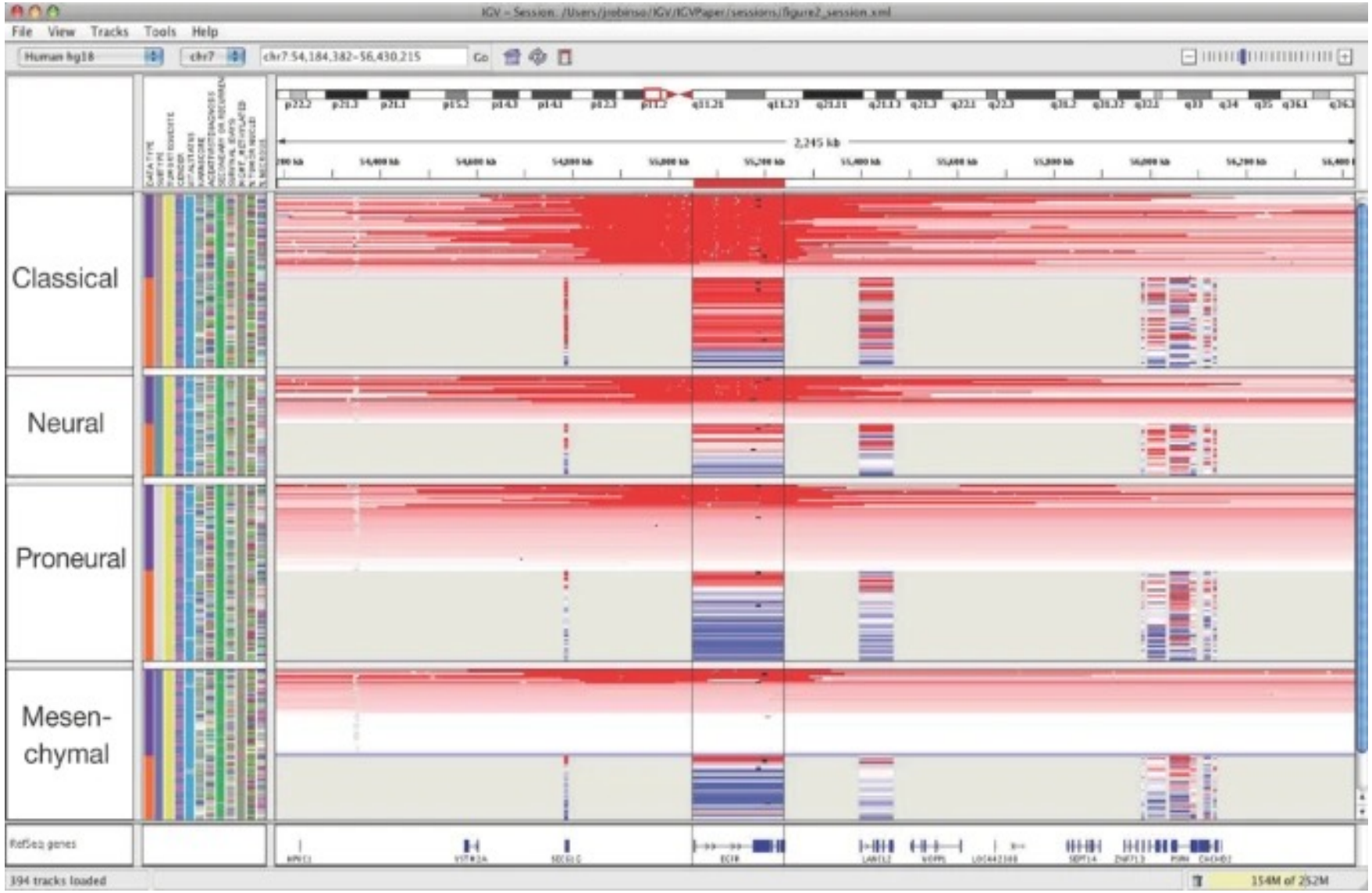
- https://cmdcolin.github.io/awesome-genome-visualization/
- Compiled by Colin Diesh
  (the lead JBrowse 2 author)

cmdcolin

UCSC XENA


UNIVERSITY OF CALIFORNIA SANTA CRUZ
DATA SETS
VISUALIZATION
TRANSCRIPTS
DATA HUBS
VIEW MY DATA
BOOKMARK
HELP
MORE TOOLS
TCGA Prostate Cancer (PRAD)
Filtered to 492 Samples
Type here or use dropper to select samples
492 matching samples
ZOOM: None
User Guide • Zoom Help • Tooltip Help
A
B
C
D
E
F
G
samples
disease
gene exp…
ERG
exon expression RNAseq - Illumi…
ERG
copy number …
TMPRSS2
copy number …
ERG
copy number segments - After re…
chr21:1-48129895
ZOOMED TO 29%
_primary_disease
samples
50 samples
prostate adenocarcinoma
100kb
ERG
5'
3'
TMPRSS2
32,873,627
1Mb
47,040,162
Click to Add Column
prostate a…
7.3
12
log2(norm_count+1)
low
high
log2(RPKM+1)
-0.93
0.93
log2(tumor/normal)
-0.39
0.39
log2(tumor/normal)
-0.39
0.39
log2(tumor/normal)

IGV

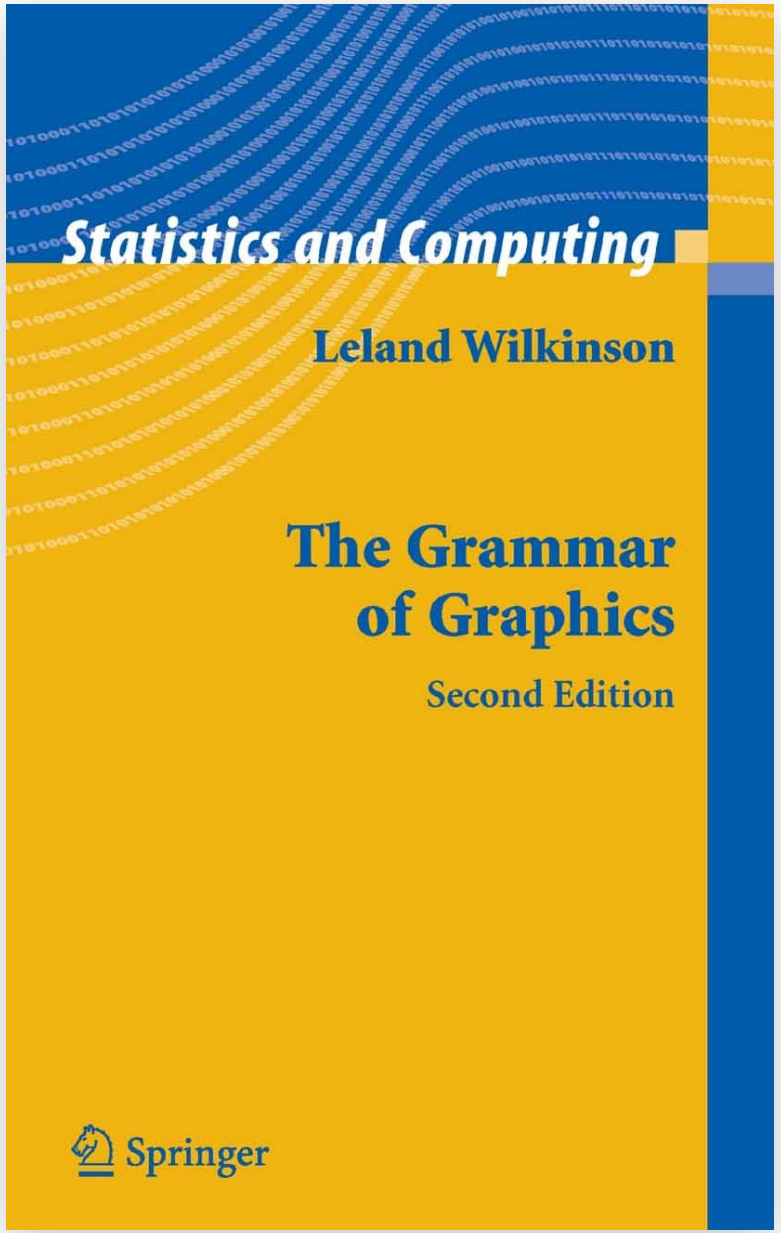

File View Tracks Tools Help
Human hg18
chr7
chr7:54,184,382-56,430,215
Go
2,245 kb
Classical
Neural
Proneural
Mesen-
chymal
RefSeq genes
394 tracks loaded
154M of 252M

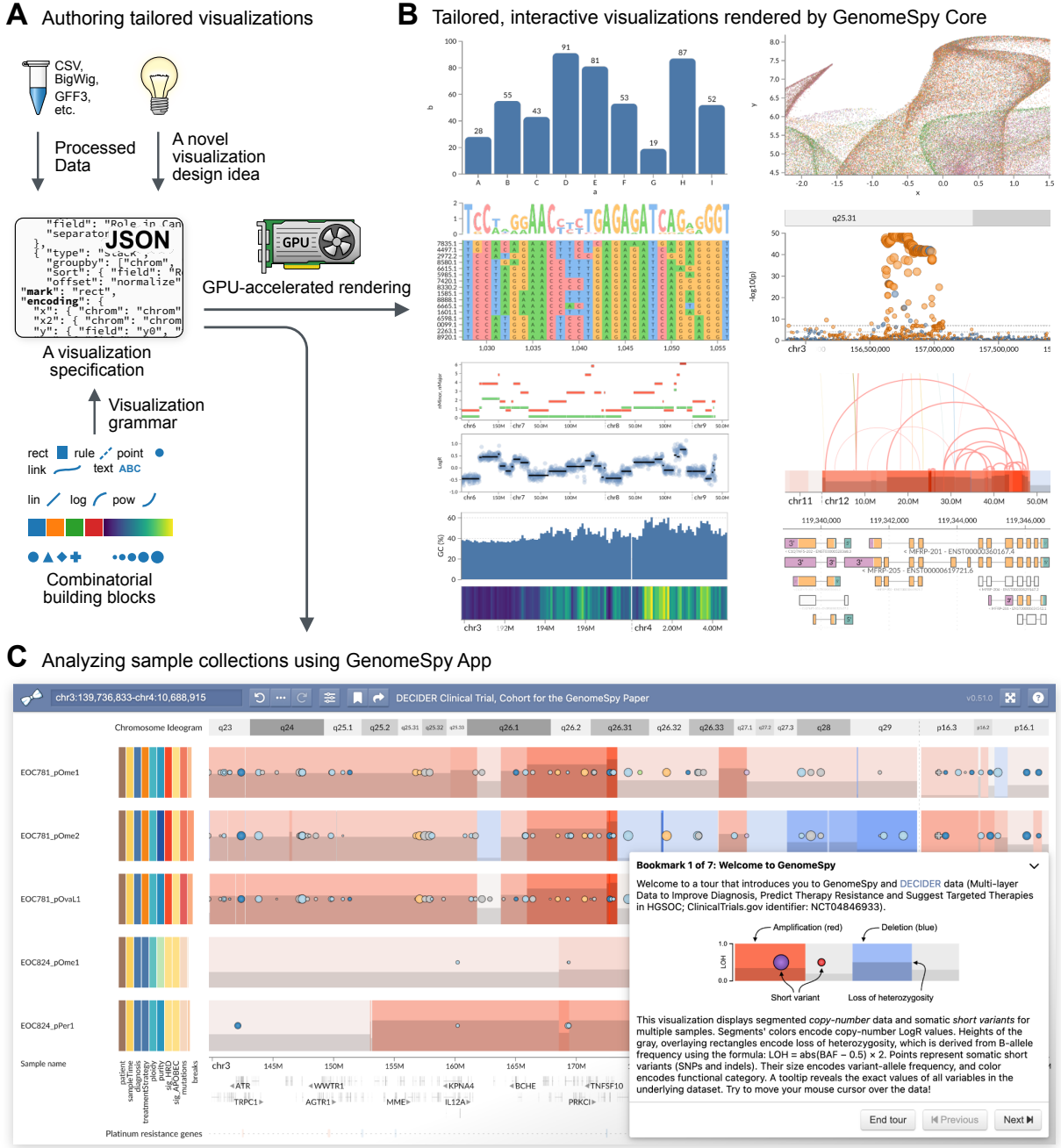
GenomeSpy

How was it born?

Genomics Game Engine?

- GPU programing is hard
- What would be a suitable abstraction?
- A Visualization Grammar!
  - Examples: ggplot2, Vega-Lite
  - More on that later…
- Developing tailored GPU-accelerated visualizations is now easy
  - At least for me 🤓




A Authoring tailored visualizations
CSV, BigWig, GFF3, etc.
Processed Data
A novel visualization design idea
JSON
A visualization specification
GPU
GPU-accelerated rendering
Visualization grammar
rect rule point
link text ABC
lin log pow
Combinatorial building blocks
B Tailored, interactive visualizations rendered by GenomeSpy Core
C Analyzing sample collections using GenomeSpy App
chr3:139,736,833-chr4:10,688,915
DECIDER Clinical Trial, Cohort for the GenomeSpy Paper
Chromosome Ideogram
EOC781_pOme1
EOC781_pOme2
EOC781_pOvaL1
EOC824_pOme1
EOC824_pPer1
Sample name
Platinum resistance genes
Bookmark 1 of 7: Welcome to GenomeSpy
Welcome to a tour that introduces you to GenomeSpy and DECIDER data (Multi-layer Data to Improve Diagnosis, Predict Therapy Resistance and Suggest Targeted Therapies in HGSOC; ClinicalTrials.gov identifier: NCT04846933).
Amplification (red)
Deletion (blue)
Short variant
Loss of heterozygosity
This visualization displays segmented copy-number data and somatic short variants for multiple samples. Segments' colors encode copy-number LogR values. Heights of the gray, overlaying rectangles encode loss of heterozygosity, which is derived from B-allele frequency using the formula: LOH = abs(BAF − 0.5) × 2. Points represent somatic short variants (SNPs and indels). Their size encodes variant-allele frequency, and color encodes functional category. A tooltip reveals the exact values of all variables in the underlying dataset. Try to move your mouse cursor over the data!
End tour
Previous
Next

Demo: SegmentModel Spy

- GATK has tons of parameters with unclear documentation. How to study their effect?
- A simple app for copy-number segmentation assessment
- Try it yourself: https://genomespy.app/segmentmodel/

Let's Explore the DECIDER Cohort

- Hautaniemi Lab geneticists uses this every day
- Navigate to: https://tinyurl.com/mr4d9tev

- Exercise: HGSC has a unique copy-number landscape. Do the copy-number peaks correlate with other peaks?

Visualization Grammars

A Brief Introduction

What Are Visualization Grammars?

- Visualization grammars are frameworks that describe how data is **transformed** into visualizations through a set of **composable rules**.
- They provide a consistent and systematic approach to creating complex visualizations by breaking them into fundamental **components**.

Wilkinson's Grammar

ELEMENT: *line*(*position*(*smooth.linear*(age*bp))
ELEMENT: *point*(*position*(age*bp))

TRANS: inverseage = *inverse*(age)
TRANS: bpbyage = *ratio*(bp, age)
ELEMENT: *line*(*position*(*smooth.linear*(inverseage*bpbyage))
ELEMENT: *point*(*position*(inverseage*bpbyage))

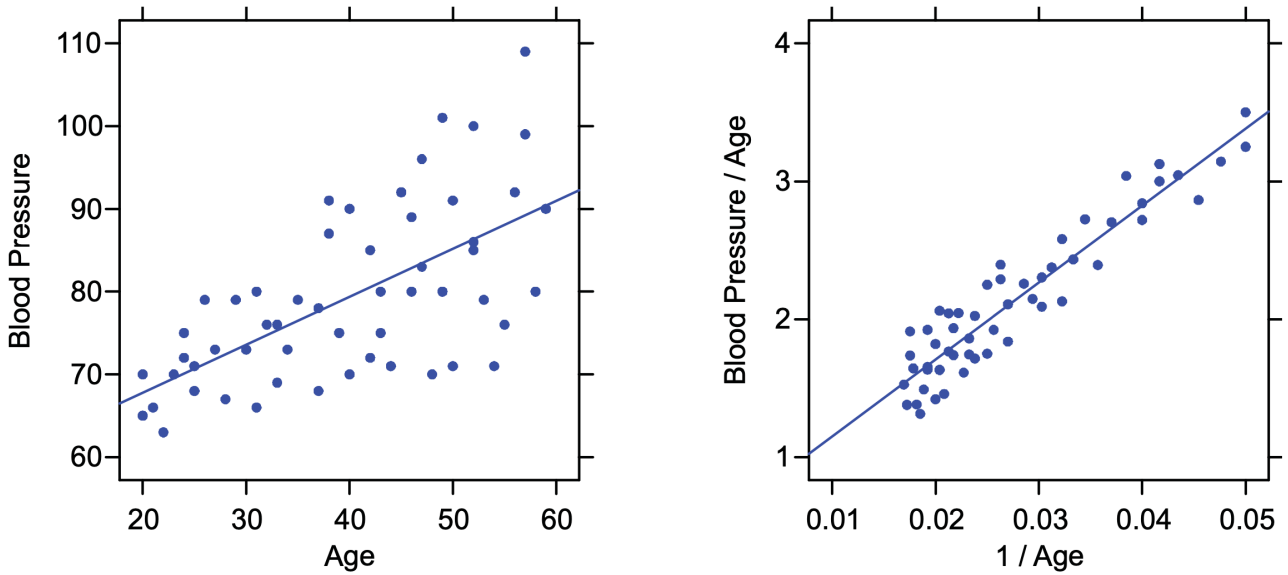
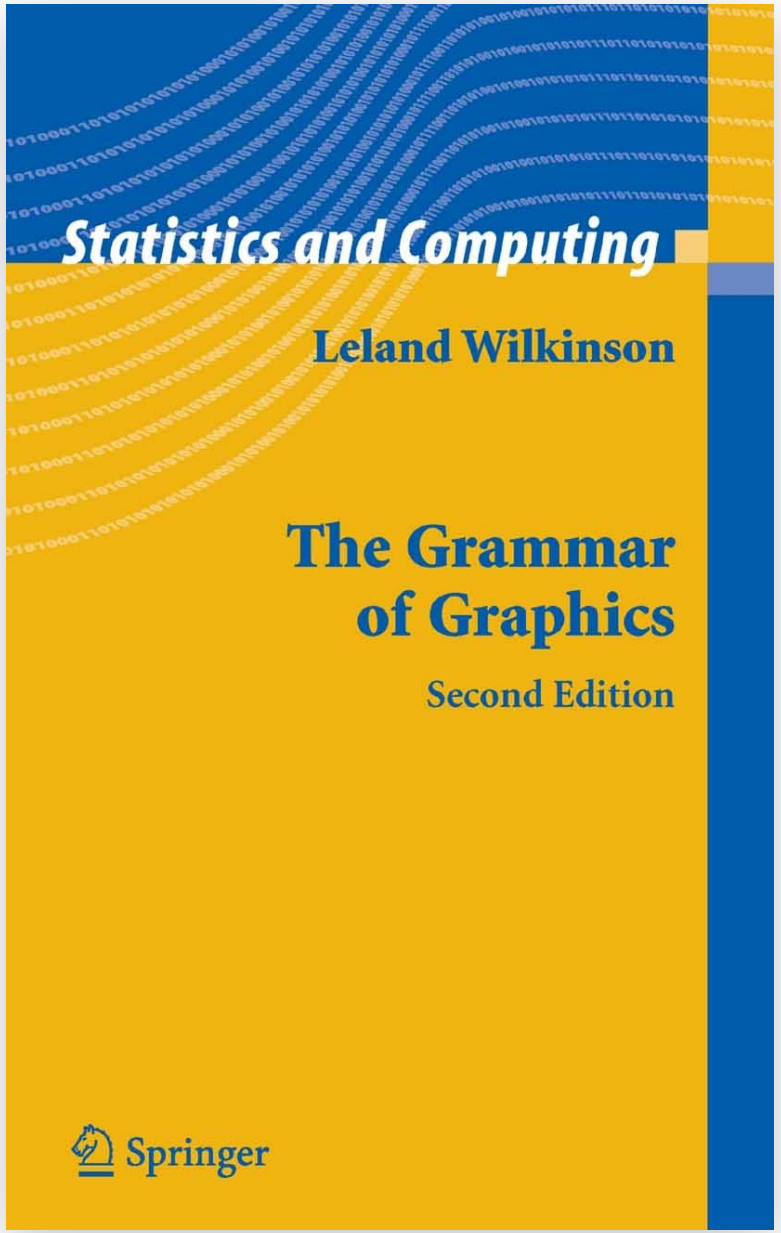
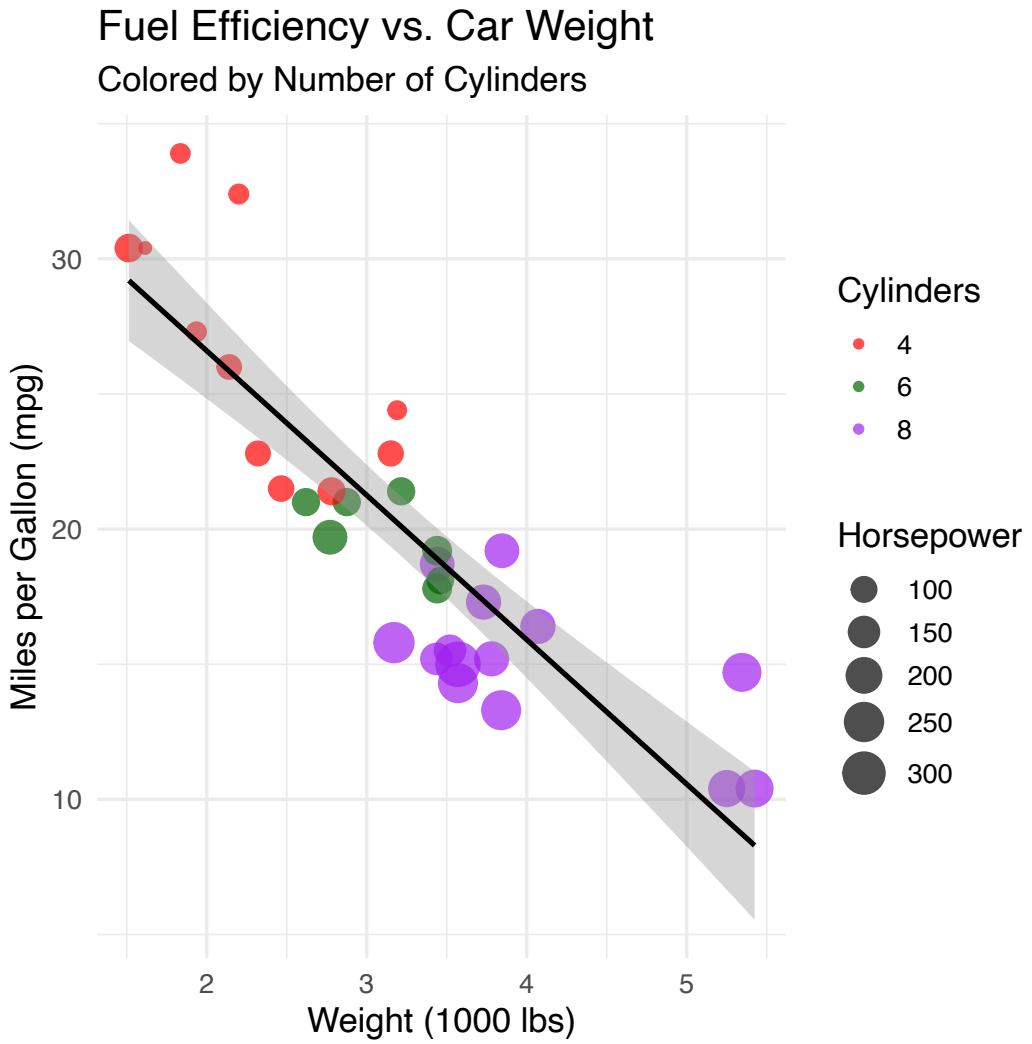


Figure 9.14 *Weighted least squares via transformation of variables*



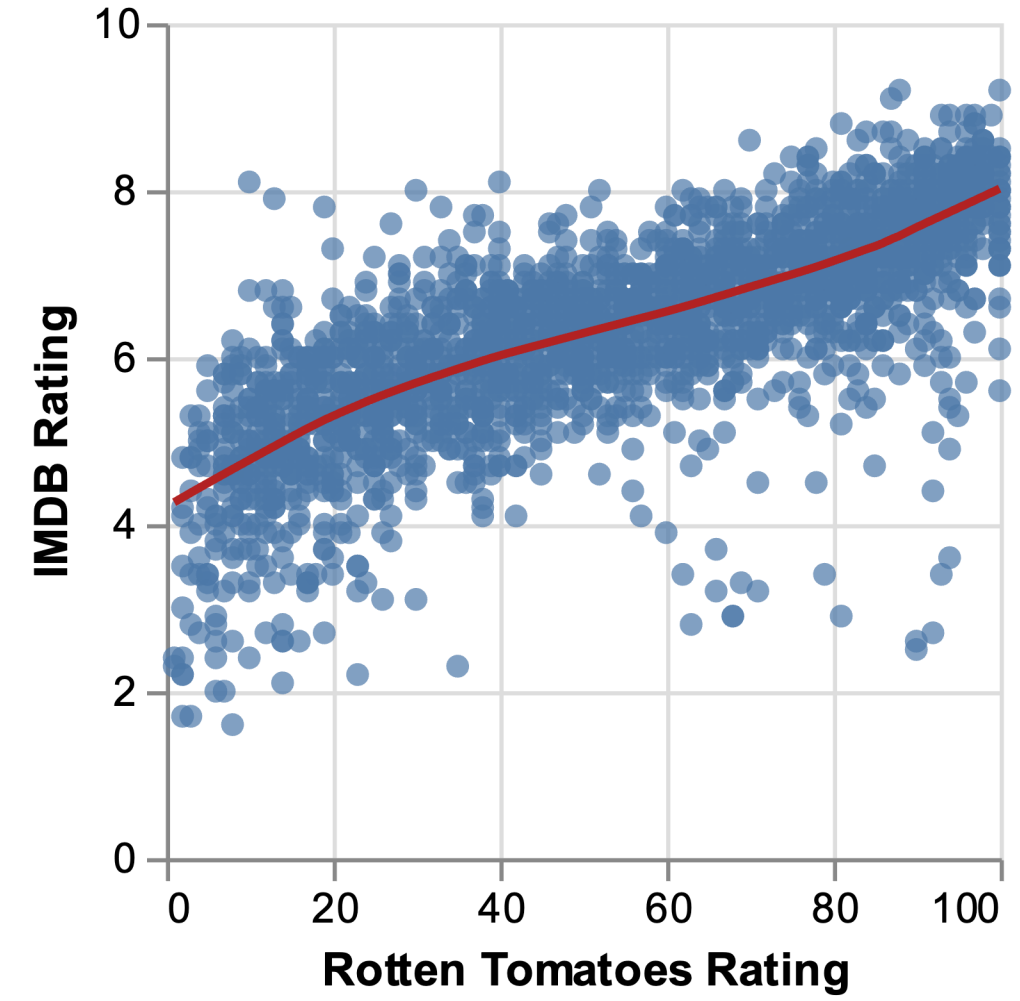
ggplot2

```
ggplot(data = mtcars, aes(x = wt, y = mpg)) +
  geom_point(aes(color = factor(cyl), size = hp),
             alpha = 0.7) +
  geom_smooth(method = "lm", color = "black") +
  scale_color_manual(values = c(
    "4" = "red",
    "6" = "darkgreen",
    "8" = "purple")) +
  scale_size_continuous(range = c(2, 8)) +
  labs(
    title = "Fuel Efficiency vs. Car Weight",
    subtitle = "Colored by Number of Cylinders",
    x = "Weight (1000 lbs)",
    y = "Miles per Gallon (mpg)",
    color = "Cylinders",
    size = "Horsepower"
  ) +
  theme_minimal(base_size = 14)
```



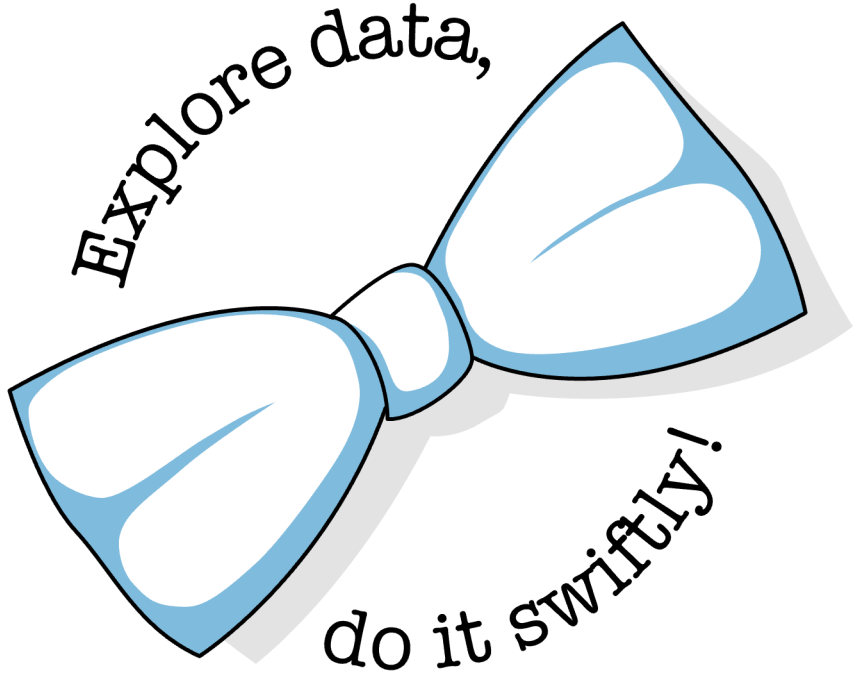
Vega-Lite

```
{
  "data": { "url": "data/movies.json" },
  "encoding": {
    "x": {
      "field": "Rotten Tomatoes Rating",
      "type": "quantitative"
    },
    "y": {
      "field": "IMDB Rating",
      "type": "quantitative"
    }
  },
  "layer": [
    {
      "mark": { "type": "point", "filled": true }
    },
    {
      "transform": [{
          "loess": "IMDB Rating",
          "on": "Rotten Tomatoes Rating"
      }],
      "mark": { "type": "line", "color": "firebrick" }
    }
  ]
}
```



GenomeSpy

- ***HEAVILY*** inspired by Vega-Lite
- Provides partial compatibility
- Genome-related enhancements
- Independent implementation
  - Scale transformations and rendering using GPU shaders
- https://genomespy.app/



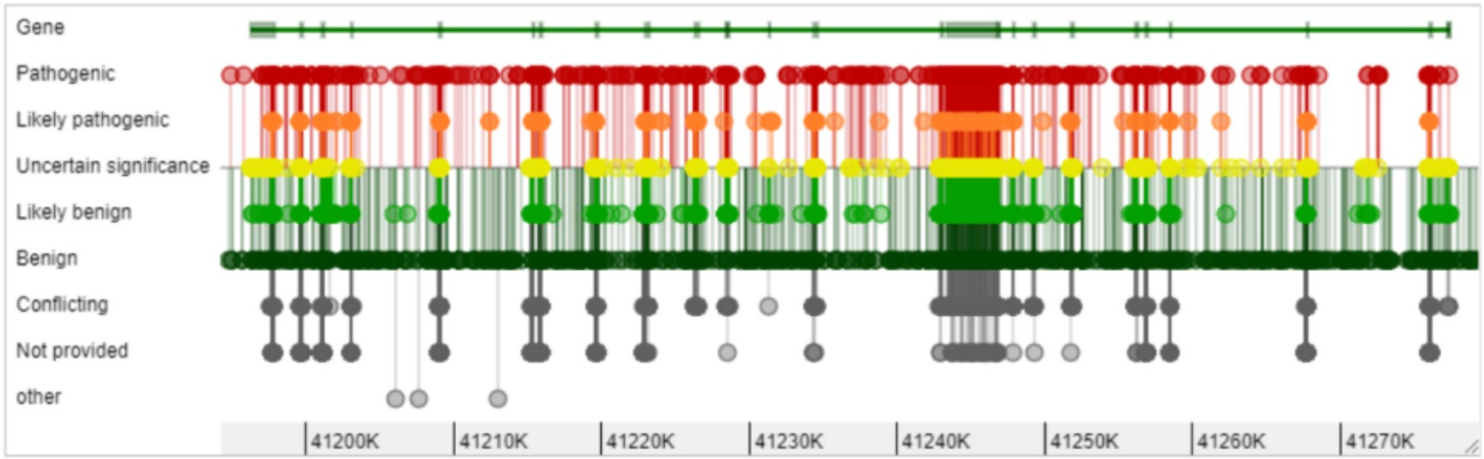
Code-Along

Let's implement a visualization!

NEW

ClinVar

New Visualization Tool



New ClinVar graphical display

Maps clinically significant variants by gene and position!

https://ncbiinsights.ncbi.nlm.nih.gov/2022/08/30/clinvar-graphical-view/

GenomeSpy Playground

- A code editor and an interactive visualization
- For small-scale testing and prototyping
- https://genomespy.app/playground/

ObservableHQ Notebooks

Embedding Visualizations

https://observablehq.com/collection/@tuner/genomespy

Local Development, Remote Deployment

Developing more complex visualizations and deploying them on remote web servers

Local Development

1. Open VS Code or another JSON-schema aware editor
2. Create an HTML file
3. Write a visualization spec
4. Start a local web server
   - `python -m http.server`
   - https://github.com/danvk/RangeHTTPServer
   - https://www.npmjs.com/package/http-server

https://genomespy.app/docs/getting-started/#local-or-remote-web-server

Remote Deployment

- No server-side logic needed
- Just put your files on a web server
  - ... and ensure that file permissions are ok

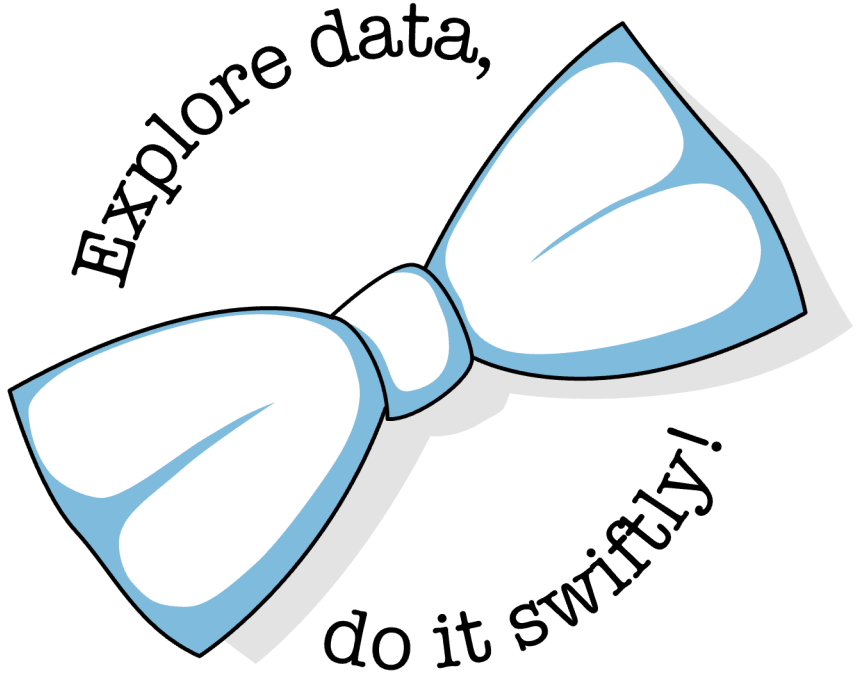
Conclusions

The Types of Visualization Challenges

- Tailored visualizations
- Large sample collections
- Finding patterns in large datasets

Future Directions

- Better developer experience
- Growing the community
- Extending the grammar
- Analysis app improvements
  - RNA expression data
  - Offer simple statistics (ANOVA, etc.)



Questions?

Follow or Contact Me for Updates

- Bluesky: @karilavikka.fi
- X: @KariLavikka
- Linkedin: https://www.linkedin.com/in/karilavikka/
- E-mail: kari.lavikka@helsinki.fi

Evaluation form