

Data Atlas for Behavior Geneticists



Wake Forest University

S. Mason Garrison





S. Mason Garrison

Hello world!



QR code for these slides

r-computing-lab.github.io/slides/00_bga_2024/d00_slide



S. Mason Garrison

The Power of Visualization

"A picture is worth a thousand words."

- + Visualizations in behavior genetics:
 - + Facilitate understanding of complex data.
 - + Reveal patterns and relationships not obvious from raw data.
 - + Enable more effective communication of research findings.

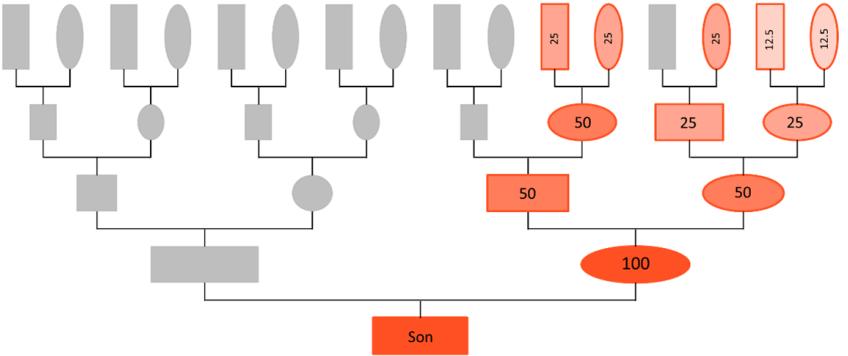


Alternative Talk

- + I *could* have spent this talk showing you some of the cool visualizations I've made since the last conference...

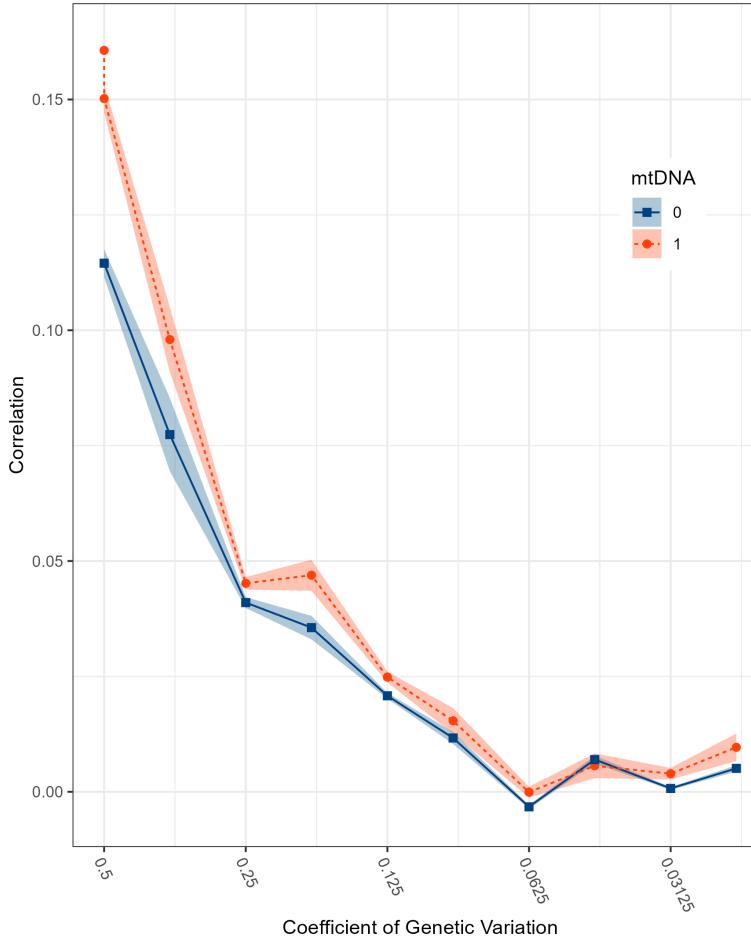


Like These Ones!



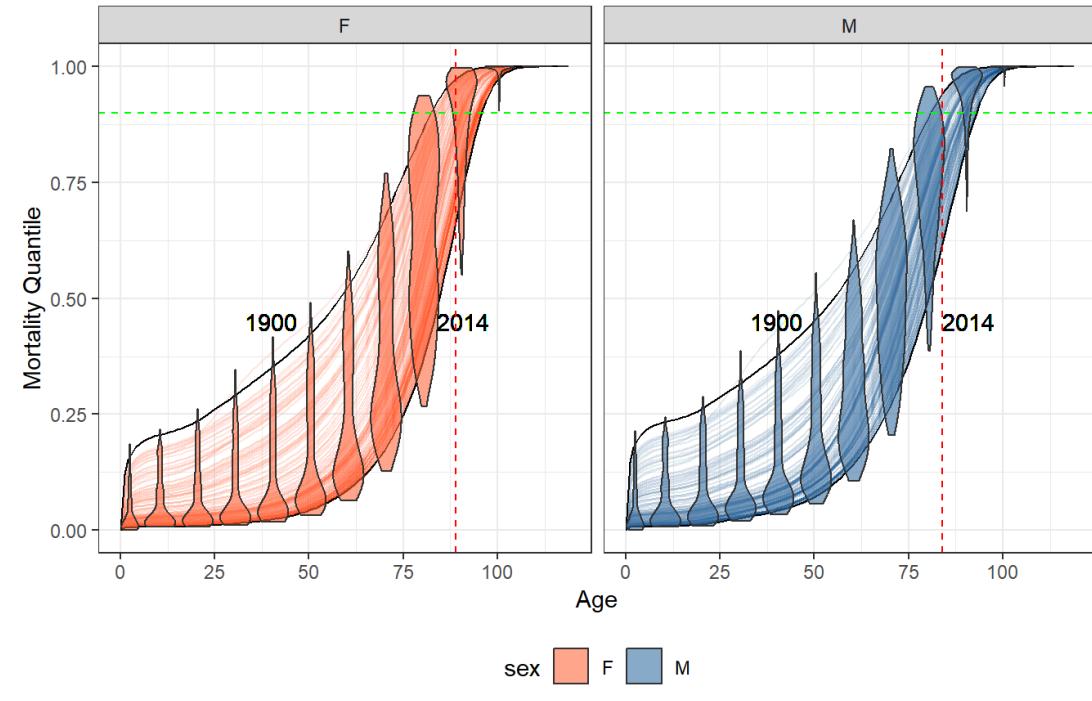
Impact of mitochondrial DNA on Longevity

Polychoric correlations between degree of relatedness
and being in the top 10% of longevity cohort adjusted using SS data.

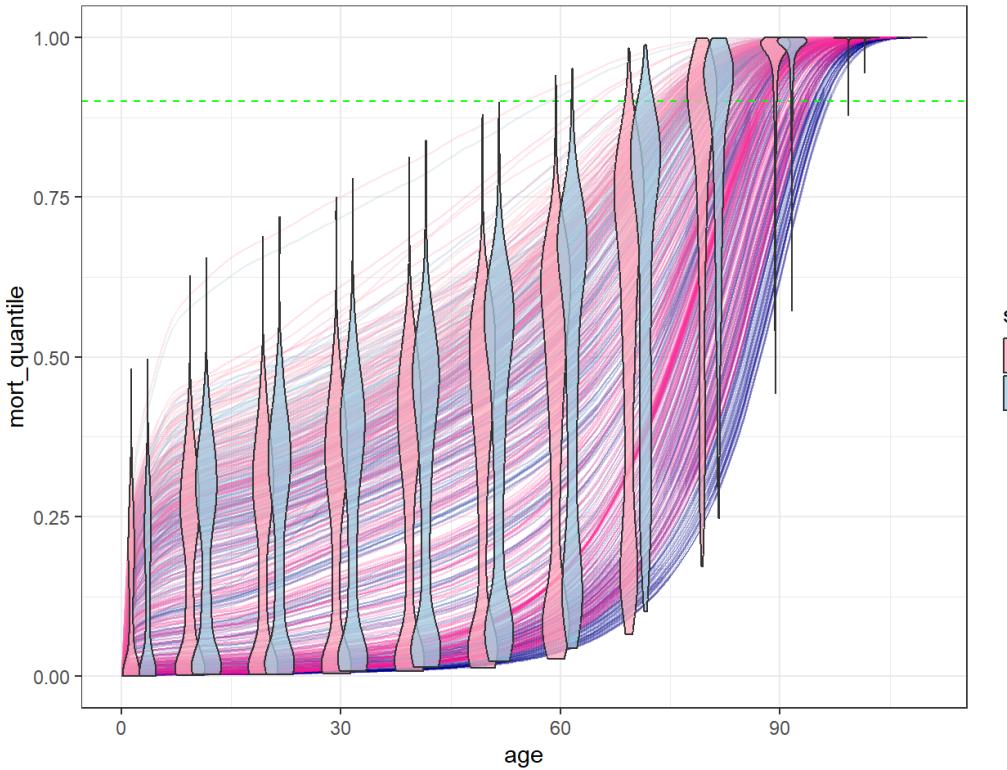


Or these!

Mortality Quantiles by Birth Cohort



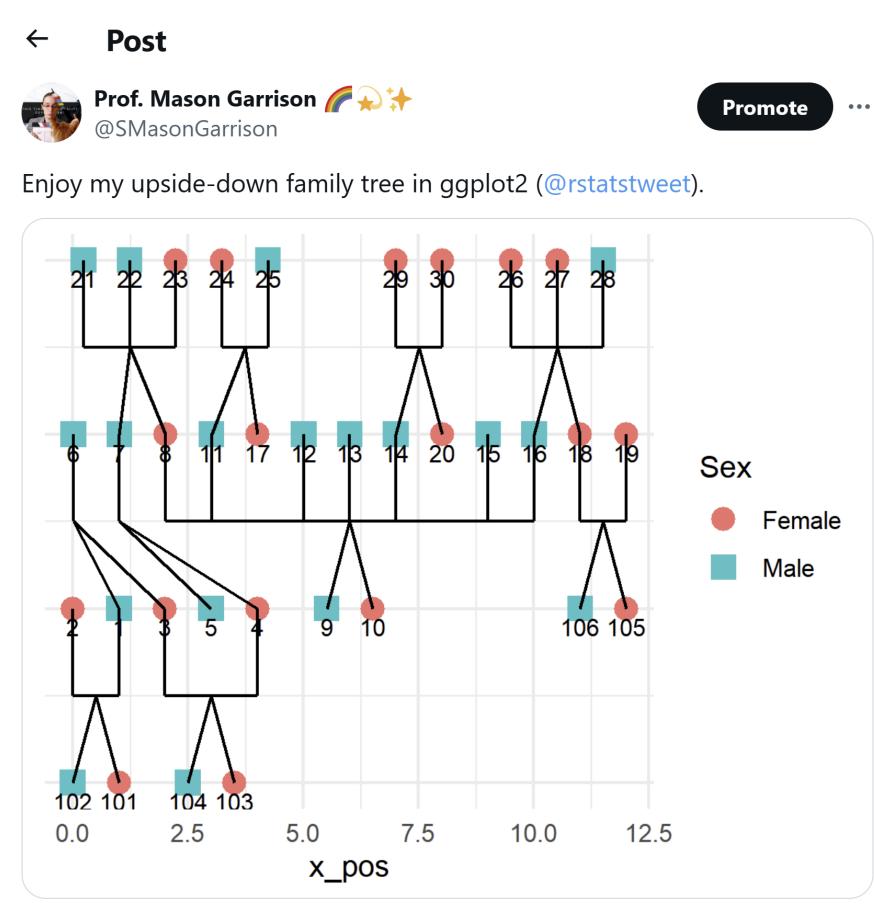
A pedigree plot of the SSA mortality from 1900 to the present.



A pedigree plot of the Swedish mortality from 1750 to the present.



Or even my latest creation!



I swear that's these are extremely promising for making much better pedigrees plots...

[x.com/SMasonGarrison/status/1800222792933310619](https://twitter.com/SMasonGarrison/status/1800222792933310619)



S. Mason Garrison

Not today!

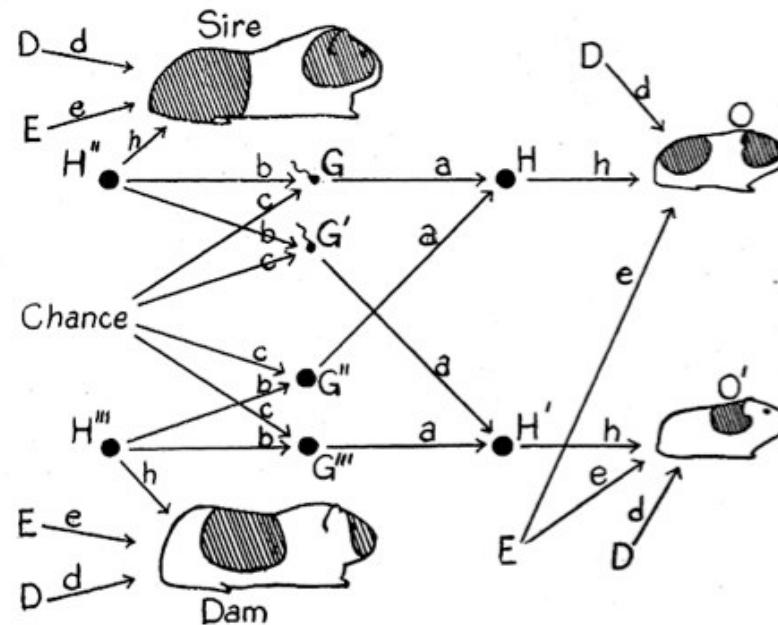


Historical Context



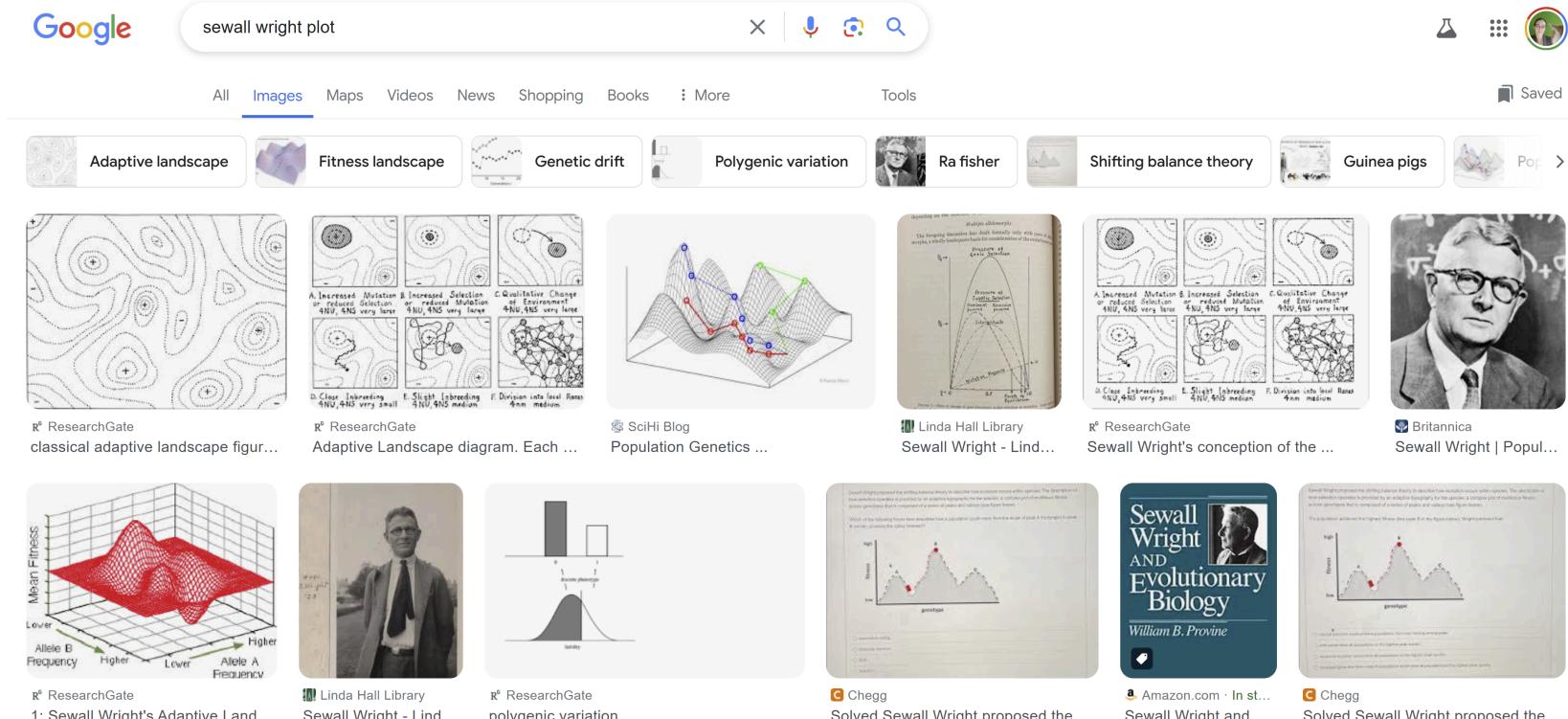
Wright's Path Diagram (1920)

- + Sewall Wright's path diagrams were pioneering tools for illustrating genetic relationships.
- + Enabled the visualization of complex genetic interactions and pathways.
- + Revolutionized genetic research by providing a clear, visual method for understanding heritability and genetic correlations.



But that's... cliché

+ Let's see some other examples from Sewall Wright's work.



Sewall Wright's 1932 Adaptive Landscape Diagram

- + Wright's adaptive landscape diagram was an attempt to reconcile Mendelian genetics and Darwinian evolution. (Shifting Balance Theory)

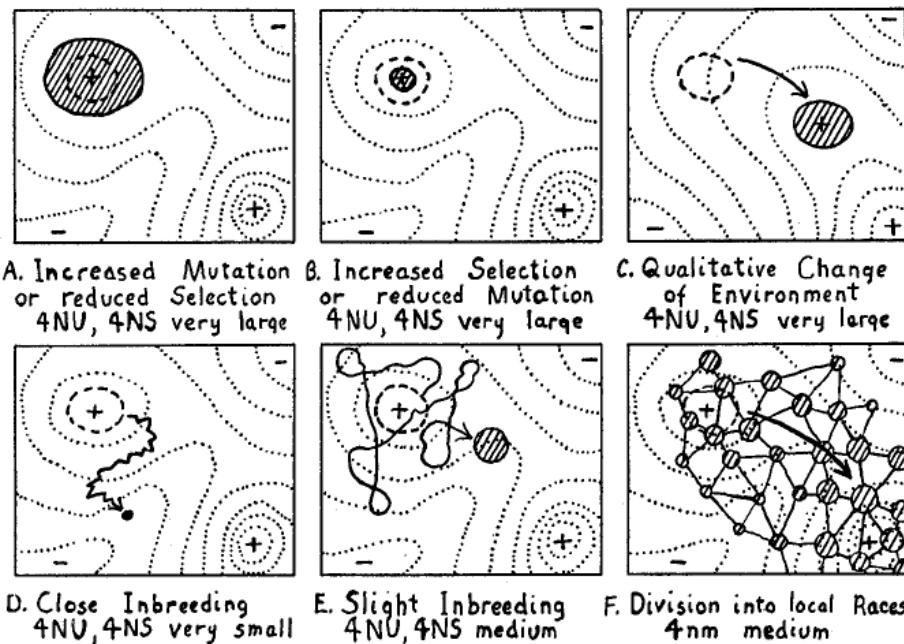


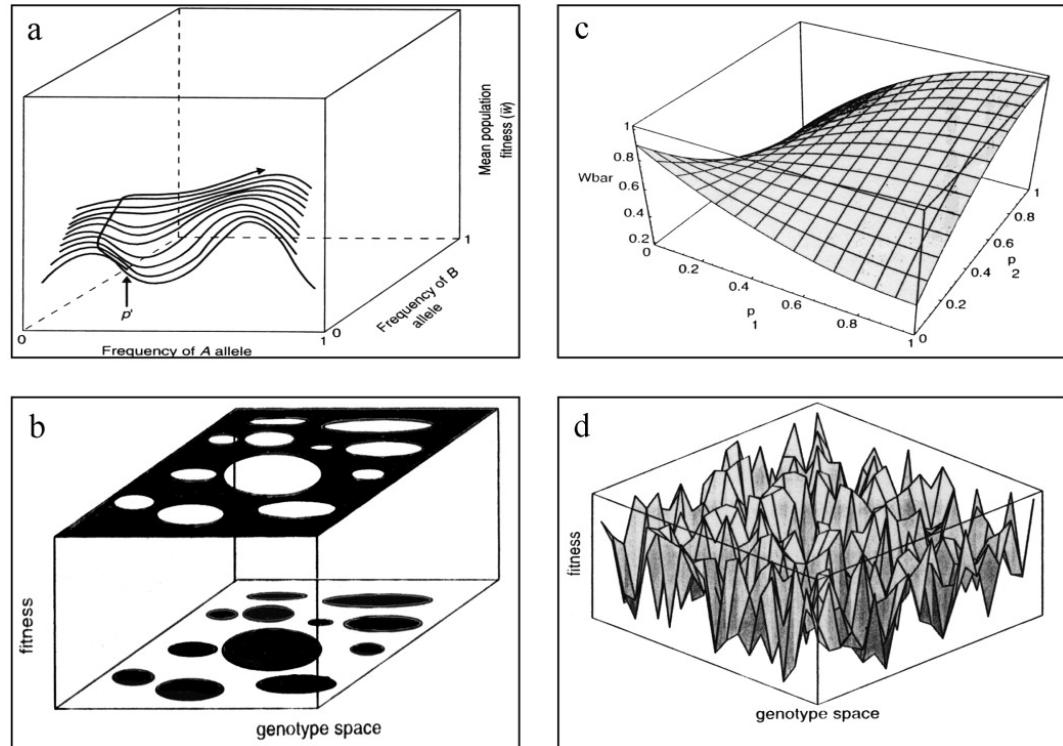
FIGURE 4.—Field of gene combinations occupied by a population within the general field of possible combinations. Type of history under specified conditions indicated by relation to initial field (heavy broken contour) and arrow.

- + Offered a metaphor of the adaptive landscape and its topographic representation.
- + Depicted the effect of variations in population size, migration, and the strength of selection.
- + Other versions of the adaptive landscape were inspired by Wright's original diagram.
 - + (Ok technically Armand Janet's conceptualized this in 1895, but Wright's concept had a much larger impact.)



Four adaptive landscape diagrams

- + Various adaptations and critiques of Wright's original diagram.
 - + A: Ridley's depiction of a Fisherian landscape (1996).
 - + B: Gavrilets' depiction of the holey landscape (1997).
 - + C: Coyne et al.'s simplified Wrightian landscape (1997).
 - + D: Kauffman and Levin's rugged adaptive landscape (1987).



Credit: Panel from Skipper (2004)



Fast Forward to the Present



S. Mason Garrison

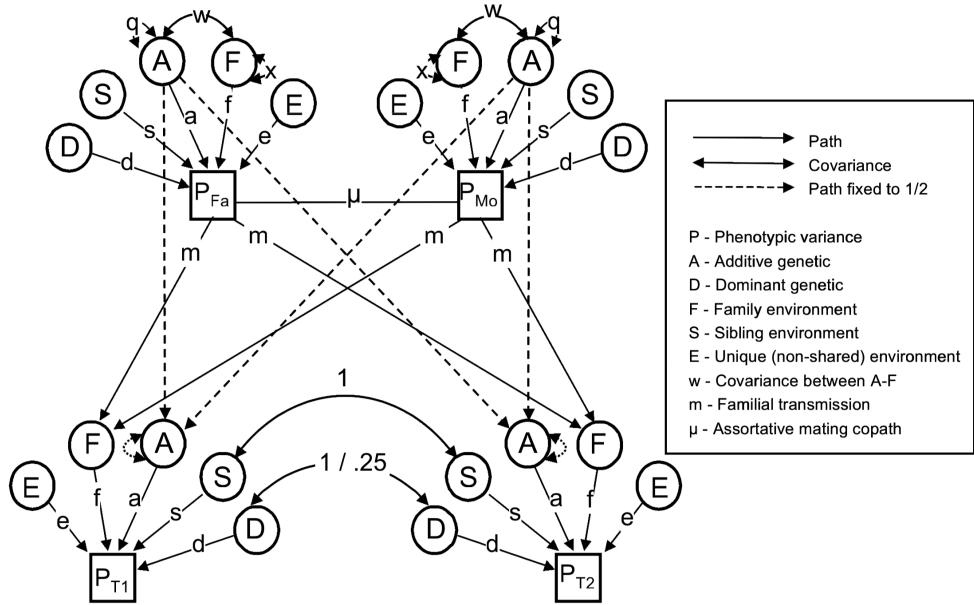
Modern Path Diagrams

- + Modern path diagrams are used in structural equation modeling (SEM).
- + Now... they use ■ s, ○ s, and ← s to represent variables, relationships, and causal pathways.
- + Visualize complex relationships between variables in a statistical model in a way that can be easily interpreted.
- + However, they can be challenging to create...



Classic path diagrams can start to resemble ...

Military Aircraft

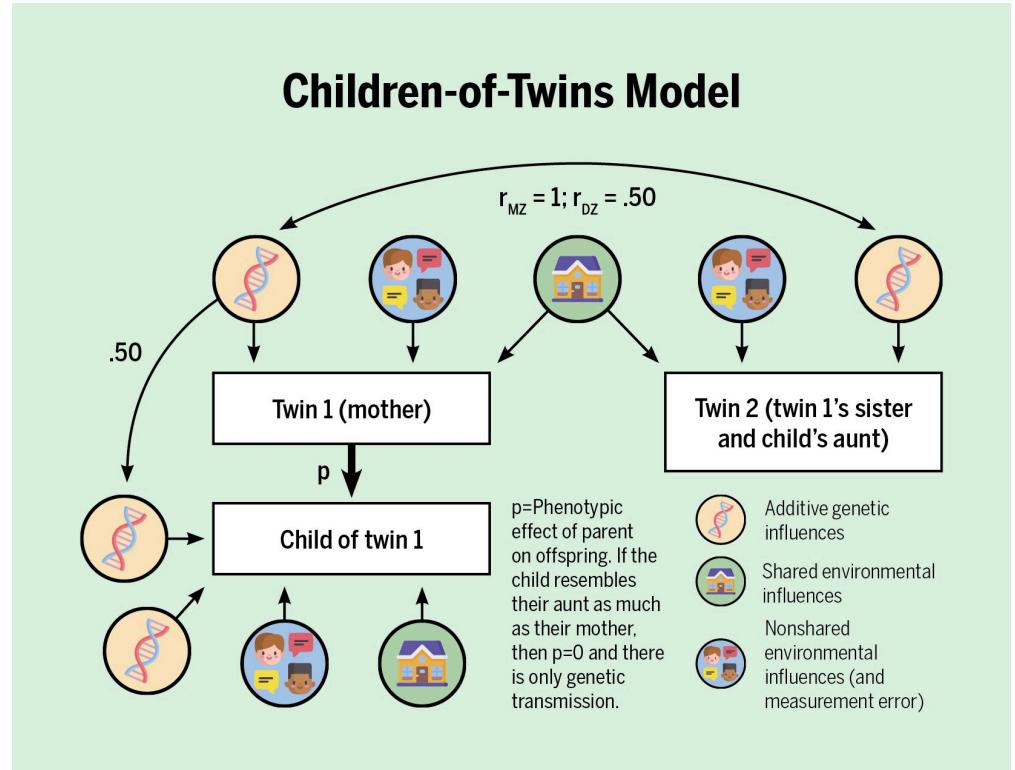


Keller, M. C., Medland, S. E., Duncan, L. E., Hatemi, P. K., Neale, M. C., Maes, H. H. M., & Eaves, L. J.

(2009). Modeling Extended Twin Family Data I: Description of the Cascade Model. *Twin Research and*

Human Genetics, 12(1), 8–18. doi:10.1375/twin.12.1.8

Storybook Characters



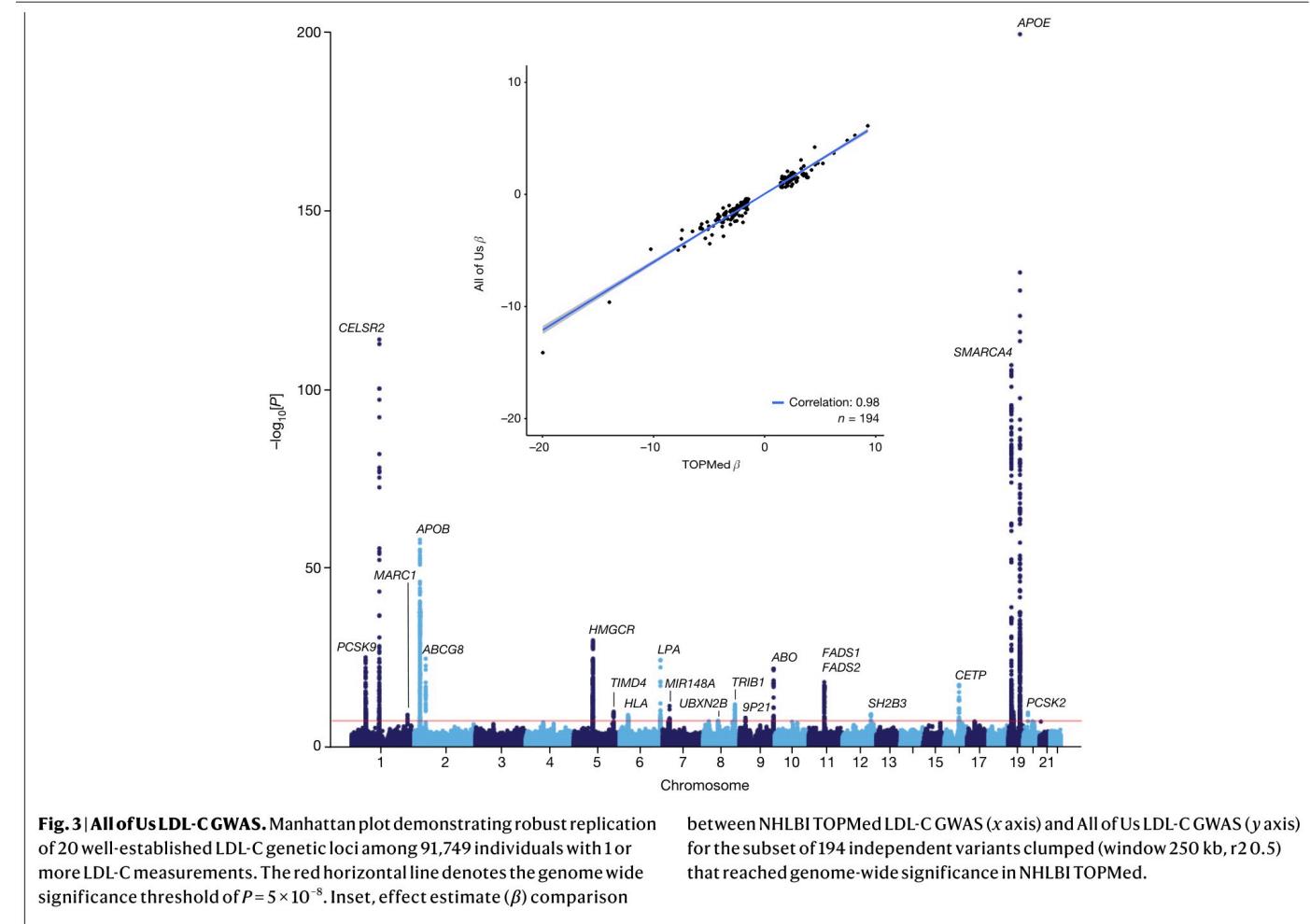
Hart, S.A., Little, C. & van Bergen, E. Nurture might be nature: cautionary tales and proposed solutions.

npj Sci. Learn. 6, 2 (2021). <https://doi.org/10.1038/s41539-020-00079-z>



Manhattan Plots

- + Widely used in genome-wide association studies (GWAS).
- + Plot p-values for genetic variants across the genome.
- + Help identify regions of the genome associated with traits or diseases.
- + Visualize large-scale genetic data, making it easier to detect significant associations.



All of Us Research Program Genomics Investigators (2024). Genomic data in the All of Us Research Program. *Nature*, 627(8003), 340–346. <https://doi.org/10.1038/s41586-023-06957-x>

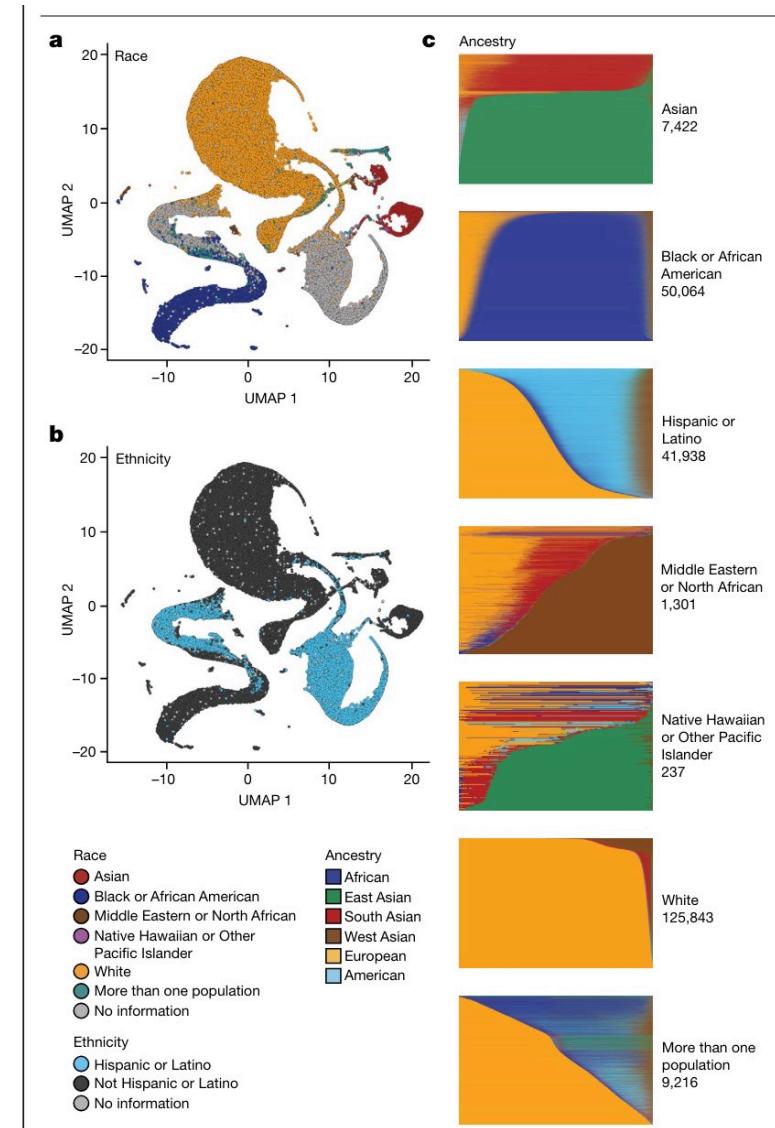


Importance of Good Graphs for Science Communication



Case Study: Misinterpretation of Genetic Data

- + A figure in a high-profile Nature paper raised concerns about reinforcing racist beliefs.
- + The UMAP generated figure aimed to show genetic diversity but was misinterpreted as supporting genetic essentialism.
- + Highlights the need for careful design and interpretation of scientific figures.

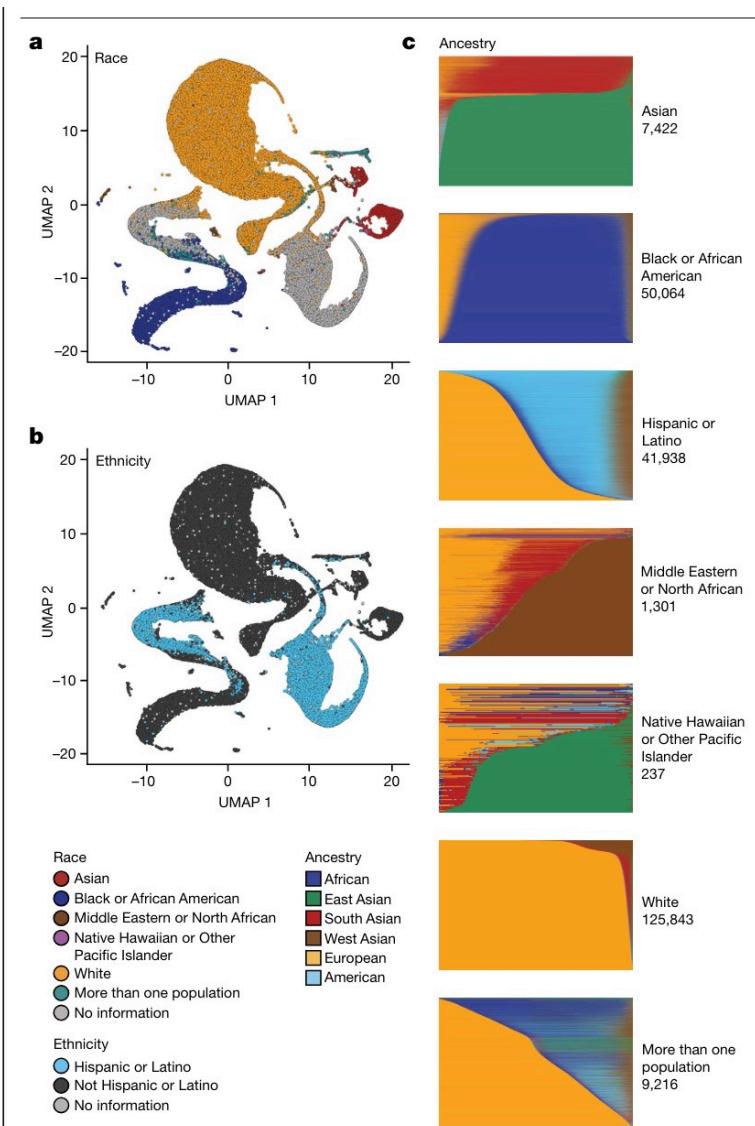
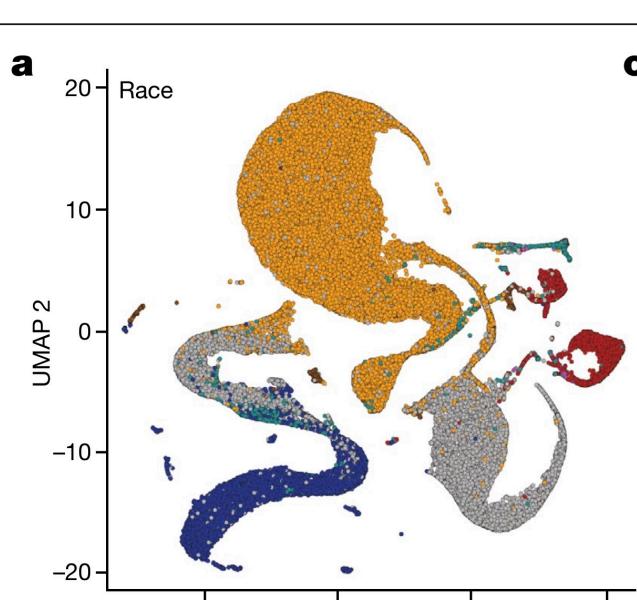


- + See more on the kerfuffle [here](#)
- + All of Us Research Program Genomics Investigators (2024). Genomic data in the All of Us Research Program. *Nature*, 627(8003), 340–346.
<https://doi.org/10.1038/s41586-023-06957-x>



Lessons Learned

- + Good graphs should accurately represent data without misleading the audience.
- + Consider the potential for misinterpretation and strive for clarity.
- + Use visualizations to enhance understanding and communication of scientific findings.



Developing Resources

Atlas with ggplot2 in R

- + A comprehensive guide for researchers to create effective visualizations using ggplot2.
- + Includes examples of various types of plots and when to use them.
- + Aims to standardize and improve the quality of visualizations in behavior genetics.
- + Provides practical tips and best practices for creating clear and informative plots.



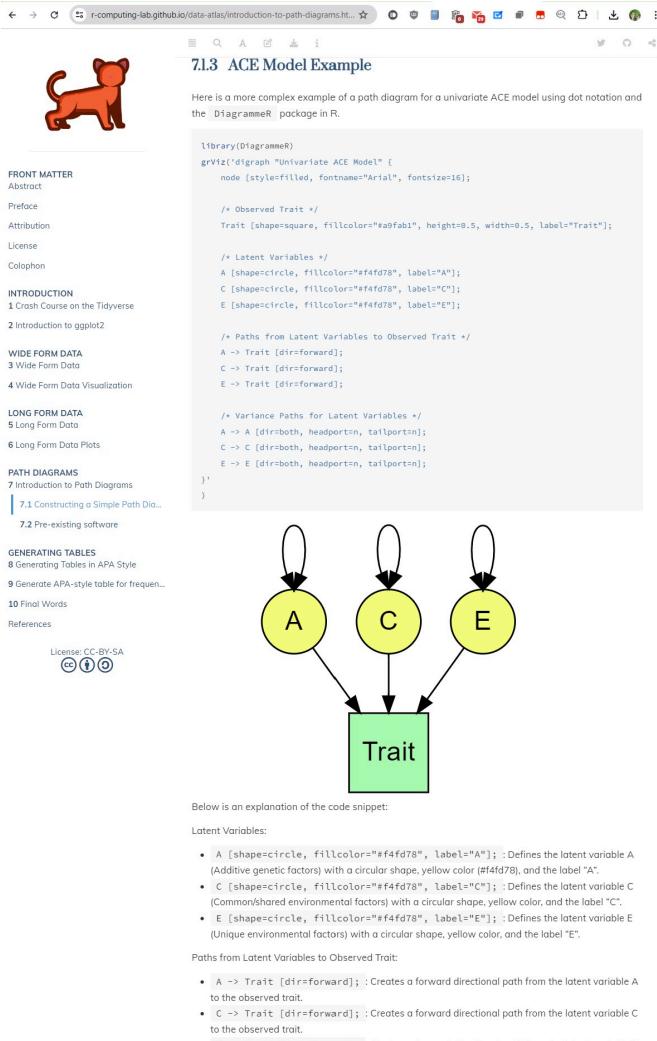
QR code for the atlas

r-computing-lab.github.io/data-atlas/



S. Mason Garrison

Atlas Path Diagrams



- + Dot Notation for specifying path diagrams.
 - + Provides a consistent and intuitive way to create path diagrams.
 - + Easy to copy and paste, and modify for different models.
 - + And easy to generate and tweak path diagrams for publication via DiagrammeR.
 - + Note: The atlas is still in very much in development and I'm happy to take requests for additional models.
 - + Note Note: Kudos to openmx and umx for their working examples of dot notation for some models



Zoom In



r-computing-lab.github.io/data-atlas/introduction-to-path-diagrams.ht... ☆

☰ ⌂ A i

7.1.3 ACE Model Example

Here is a more complex example of a path diagram for a univariate ACE model using dot notation and the `DiagrammeR` package in R.

```
library(DiagrammeR)
grViz('digraph "Univariate ACE Model" {
    node [style=filled, fontname="Arial", fontsize=16];

    /* Observed Trait */
    Trait [shape=square, fillcolor="#a9fab1", height=0.5, width=0.5, label="Trait"];

    /* Latent Variables */
    A [shape=circle, fillcolor="#f4fd78", label="A"];
    C [shape=circle, fillcolor="#f4fd78", label="C"];
    E [shape=circle, fillcolor="#f4fd78", label="E"];

    /* Paths from Latent Variables to Observed Trait */
    A -> Trait [dir=forward];
    C -> Trait [dir=forward];
    E -> Trait [dir=forward];

    /* Variance Paths for Latent Variables */
    A -> A [dir=both, headport=n, tailport=n];
    C -> C [dir=both, headport=n, tailport=n];
}
```

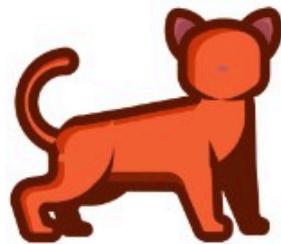


Atlas Correlations

← → C r-computing-lab.github.io/data-atlas/wide-form-data-visualization.html ☆ ⚡ 🛡️ 📁 📤 📺 📲 📱 🔍 🗂️ 📲 🐾 🌐 📲 ⋮

☰ 🔎 A 🖊️ 📽️ i

🐦 🌀 🏷️ 🖼️



4.3.1 Correlation Matrix and Correlogram

```
library(ggcorrplot)

# select only the variables of interest

df_cor <- df_wide %>% select(wt1, wt2, ht1, ht2)

# Compute correlation matrix

corr <- cor(df_cor ,use="pairwise.complete") %>% round(2)

ggcorrplot(corr, type = "lower", lab = TRUE,
           lab_size = 3,
           method = "circle",
           colors = c("tomato2", "white", "springgreen3"),
           title = "Correlation Matrix of Twin Data",
           ggtheme = theme_bw)
```

FRONT MATTER

Abstract

Preface

Attribution

License

Colophon

INTRODUCTION

1 Crash Course on the Tidyverse

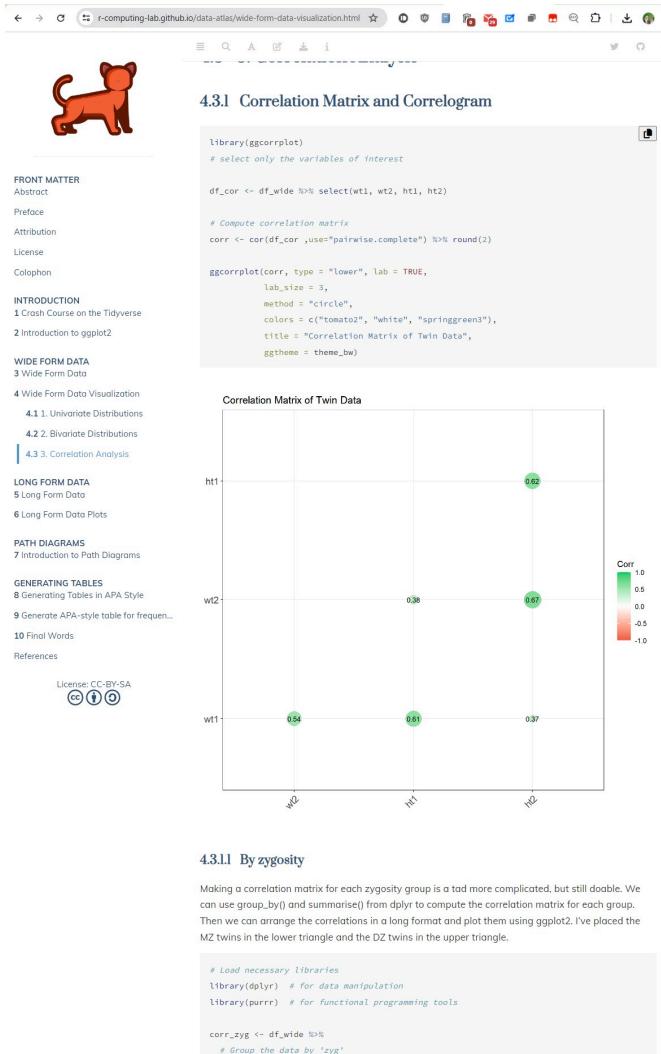
2 Introduction to ggplot2



WIDE FORM DATA

S. Mason Garrison

Zoom out



The screenshot shows a web browser displaying a page from r-computing-lab.github.io/data-atlas/wide-form-data-visualization.html. The page title is "4.3.1 Correlation Matrix and Correlogram".

The left sidebar contains a navigation menu with sections like FRONT MATTER, INTRODUCTION, WIDE FORM DATA, PATH DIAGRAMS, GENERATING TABLES, and REFERENCES. A small orange cat icon is at the top of the sidebar.

The main content area shows R code for generating a correlation matrix and a correlogram, followed by a "Correlation Matrix of Twin Data" plot. The plot shows correlations between variables wt1, ht1, wt2, and ht2. The color scale ranges from -1.0 (red) to 1.0 (green). The correlation values are:

Variable	wt1	ht1	wt2	ht2
wt1	0.54	0.61	0.37	0.62
ht1	0.62	0.38	0.67	
wt2	0.37	0.67	0.62	
ht2				0.62

Below the plot, there is a section titled "4.3.1.1 By zygosity" with a note about creating a correlation matrix for zygosity groups.

The bottom of the page includes a footer with the text "S. Mason Garrison" and the orange cat icon.

```
library(ggcorrplot)
# select only the variables of interest
df_cor <- df_wide %>% select(wt1, wt2, ht1, ht2)

# Compute correlation matrix
corr <- cor(df_cor, use="pairwise.complete") %>% round(2)

ggcorrplot(corr, type = "lower", lab = TRUE,
           lab_size = 3,
           method = "circle",
           colors = c("tomato2", "white", "springgreen3"),
           title = "Correlation Matrix of Twin Data",
           ggtheme = theme_bw)
```

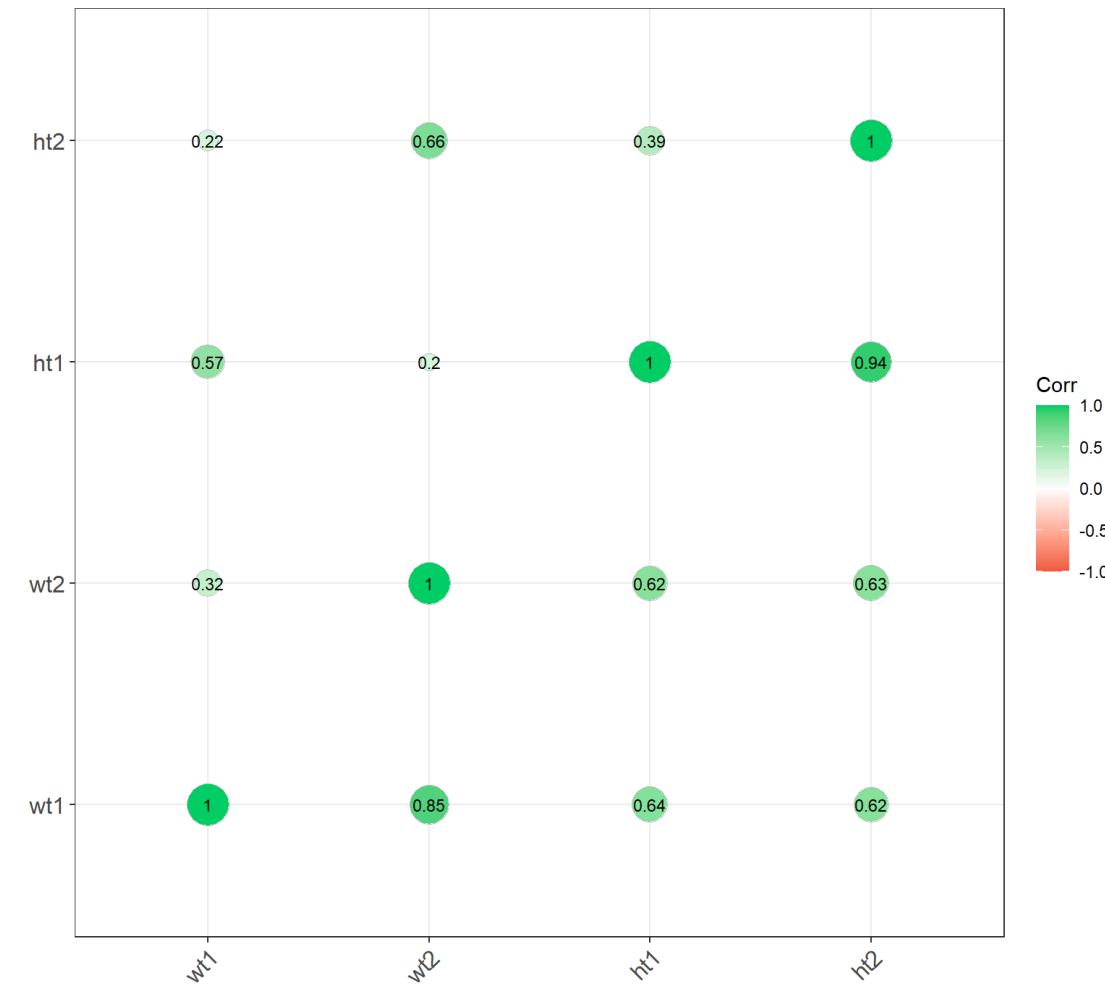
```
# Load necessary libraries
library(dplyr) # for data manipulation
library(purrr) # for functional programming tools

corr_zyg <- df_wide %>%
  # Group the data by "zyg"
```

Zoom less out

The screenshot shows a web-based R environment. At the top, there's a navigation bar with icons for back, forward, search, and file operations. Below the bar is a header with a small orange cat icon and the URL "r-computing-lab.github.io/data-atlas/wide-form-data-visualization.html". The main content area has a title "4.3.1 Correlation Matrix and Correlogram" and a sub-section "4.3.3. Correlation Analysis". It contains R code for generating a correlation matrix and a correlogram, followed by a "Correlation Matrix of Twin Data" plot. The plot is a lower triangular matrix with variables wt1, ht1, wt2, and ht2 on both axes. The diagonal elements are 1.0, and the off-diagonal elements show correlations: wt1-wt2 is 0.38, ht1-wt2 is 0.67, ht1-ht2 is 0.62, wt1-ht2 is 0.37, and wt1-ht1 is 0.64. A color scale legend on the right indicates correlation values from -1.0 (red) to 1.0 (green). Below the plot, there's a section titled "4.3.1.1 By zygosity" with a note about computing correlations for different zygosity groups.

Correlation Matrix of Twin Data by Zygosity



MZ correlations in the lower triangle,
DZ correlations in the upper triangle



Proposed R Package ?

Integration with OpenMx:

- + Streamlines the creation of visualizations for structural equation modeling (SEM) and other complex analyses. (Probably attempting to generalize DiagrammeR to works beyond RAM models)
- + Provides functions for creating path diagrams, correlation matrices, and other visualizations compatible with the data structures used in OpenMx.
- + Aims to make advanced visualization tools more accessible to researchers in behavior genetics.



Democratizing Visualization Tools

Making Advanced Scripts Accessible



QR code for the atlas

- + Open-source resources: Free and available to the entire research community.
- + User-friendly packages: Simplify the process of creating high-quality visualizations.
- + Educational materials: Provide guidance and training for researchers at all levels. Designed to complement what is already available from IBG and openMx and my own Data Science Materials.

Benefits:

- + Improves the quality of visualizations in behavior genetics.
- + Increases the impact and reach of research findings.
- + Amplify the reach and impact of behavior genetics research.
- + Support the behavior genetics community with more tools and resources.



Conclusion and Q&A

Advancing the Field through Visualization

- + Today, we delved into some of the classic data visualizations in behavior genetics.
- + We traced its evolution from Wright's path diagrams to modern Manhattan plots.
- + We highlighted the development of new resources to enhance visualization practices in the field.
- + Remember, effective visualization can transform complex data into understandable insights. Embrace these tools and improve your research communication.

Thank you for your attention.



Any Questions?

Feel free to ask any questions now, or reach out to me after the talk via email garrissm@wfu.edu or on github github.com/smasongarrison.



r-computing-lab.github.io/data-atlas/



S. Mason Garrison