Data Visualization

Sarah Pendergrass, PhD MS
Research Associate
Center for Systems Genomics

Outline

- Why data visualization
  - Examples of the power of data visualization
  - A few rules of thumb
  - Breaking out of tables
- Data visualization software

Remember: these methods scale up to "big data"!
What are the things you want to plot/visualize?

Why Data Visualization?

- Large amount of complex data
  - GWAS results
  - Multiple data types
  - Emerging high throughput sequencing methods

Why Data Visualization?

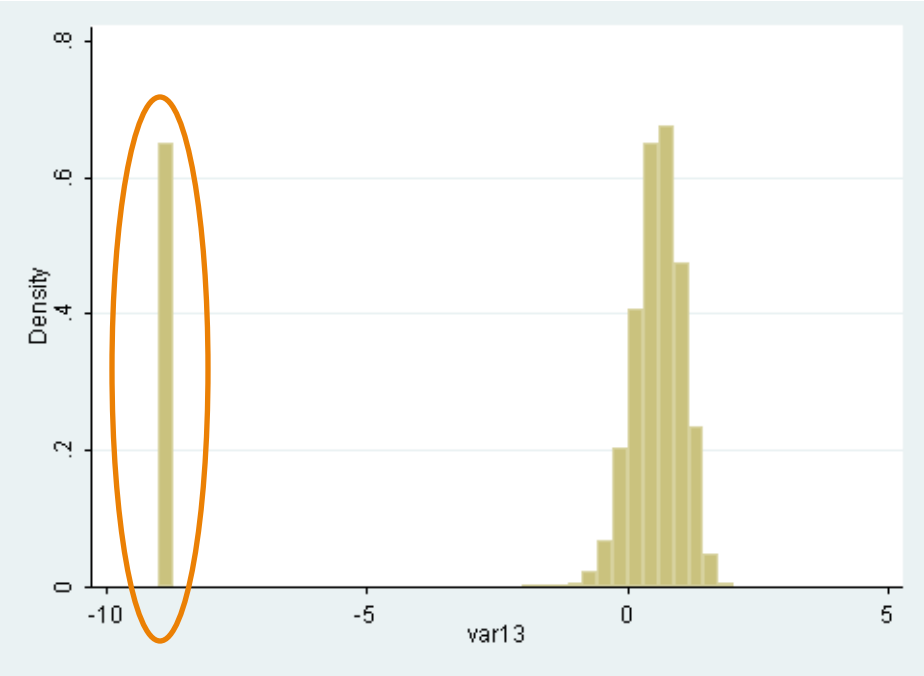
- Visualizing data
  - Highlighting issues or areas of further investigation

    - Important for all stages of analysis
      - Understanding your data
      - The "sense-making" process
        - Asking questions, exploration

      - Sharing results
        - Discussions
        - Presentations
        - Publications

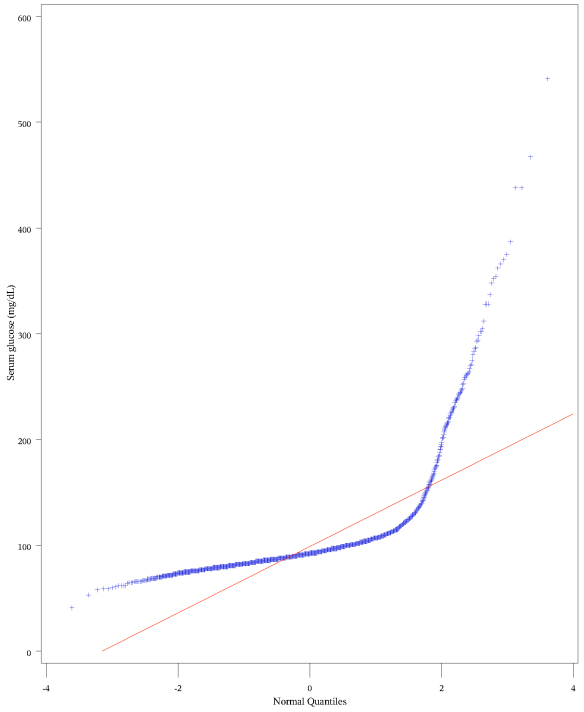
Landscape of the Data

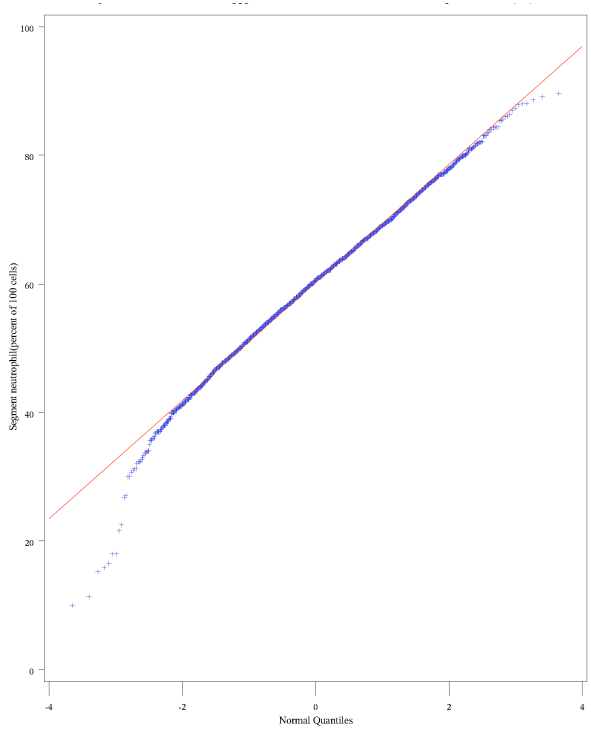
- Understanding your data
- Phenotype inspection
  - Potential problems or need for transformation
- Trends in measurements



Landscape of the Data

- Phenotype inspection
  - Potential problems or need for transformation
- Trends in measurements



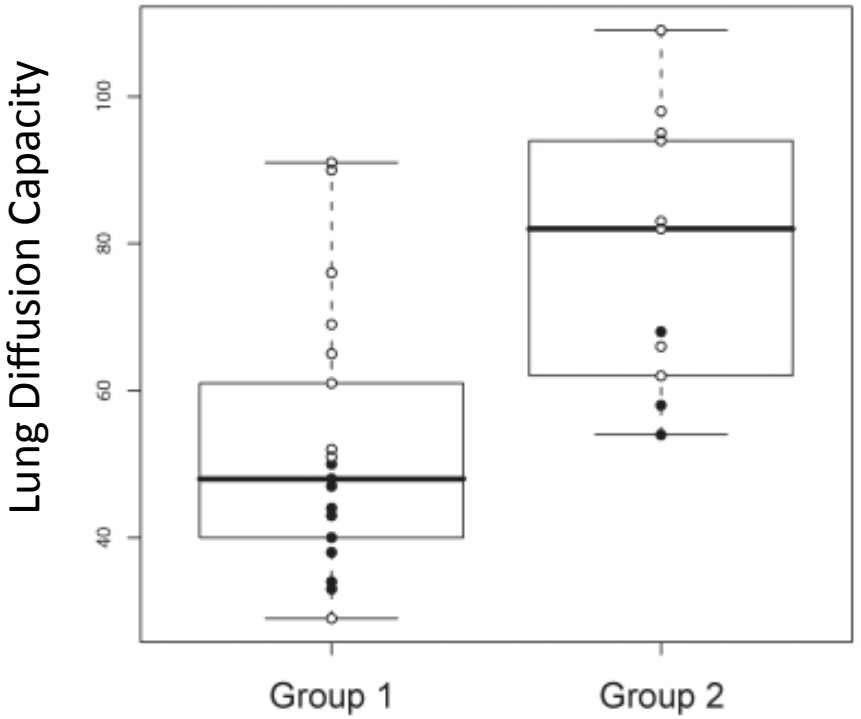


Landscape of the Data

- Phenotype inspection
  - Potential problems
- Trends in measurements

Open Circle: No PAH
Closed Circle: PAH

Lung Diffusion Capacity
Differences Between Groups



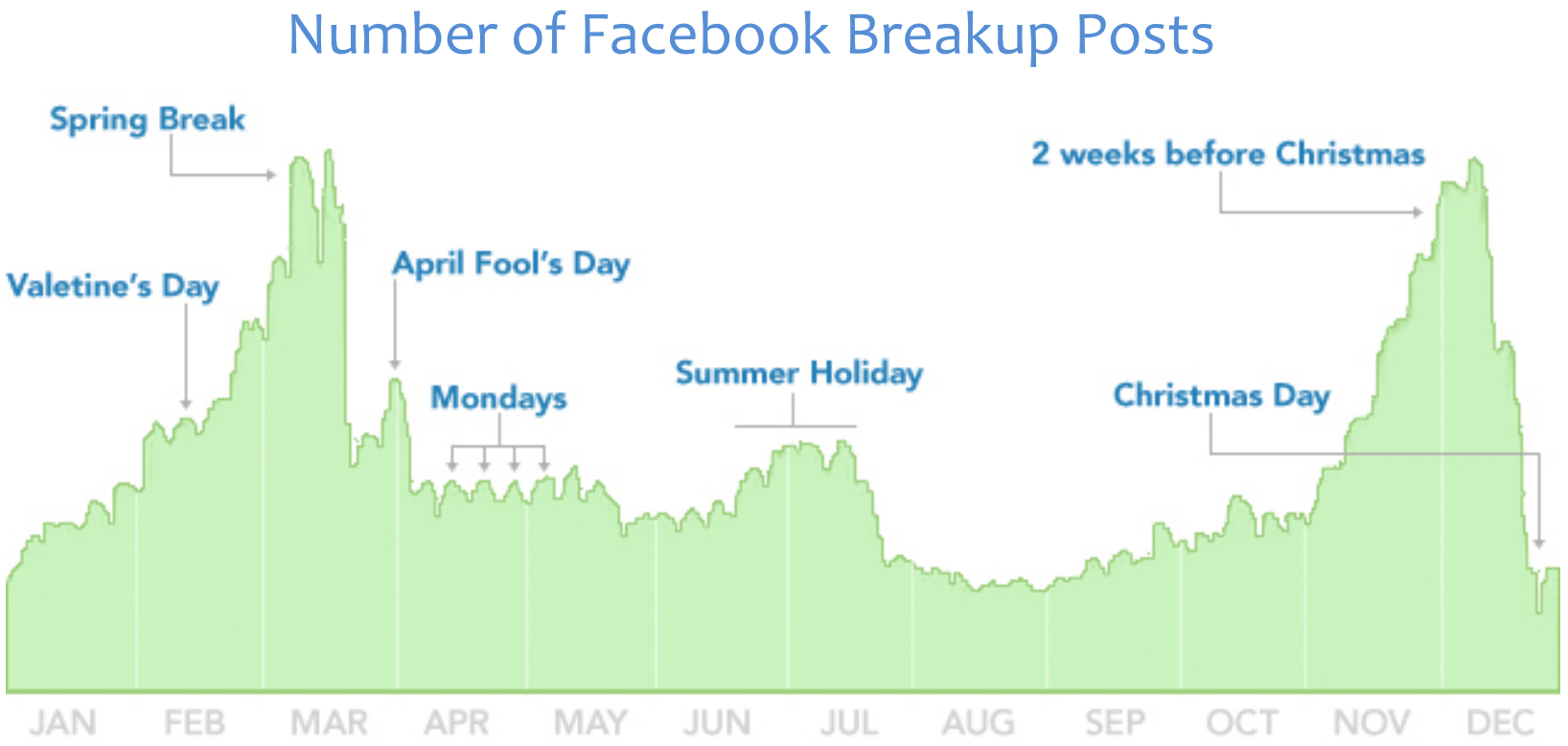
Graphic: Pendergrass et al. 2010

Why Data Visualization?

- Visualizing data
  - Highlighting issues or areas of further investigation
    - Important for all stages of analysis
  - Expose patterns and connections within and between data

Why Data Visualization?

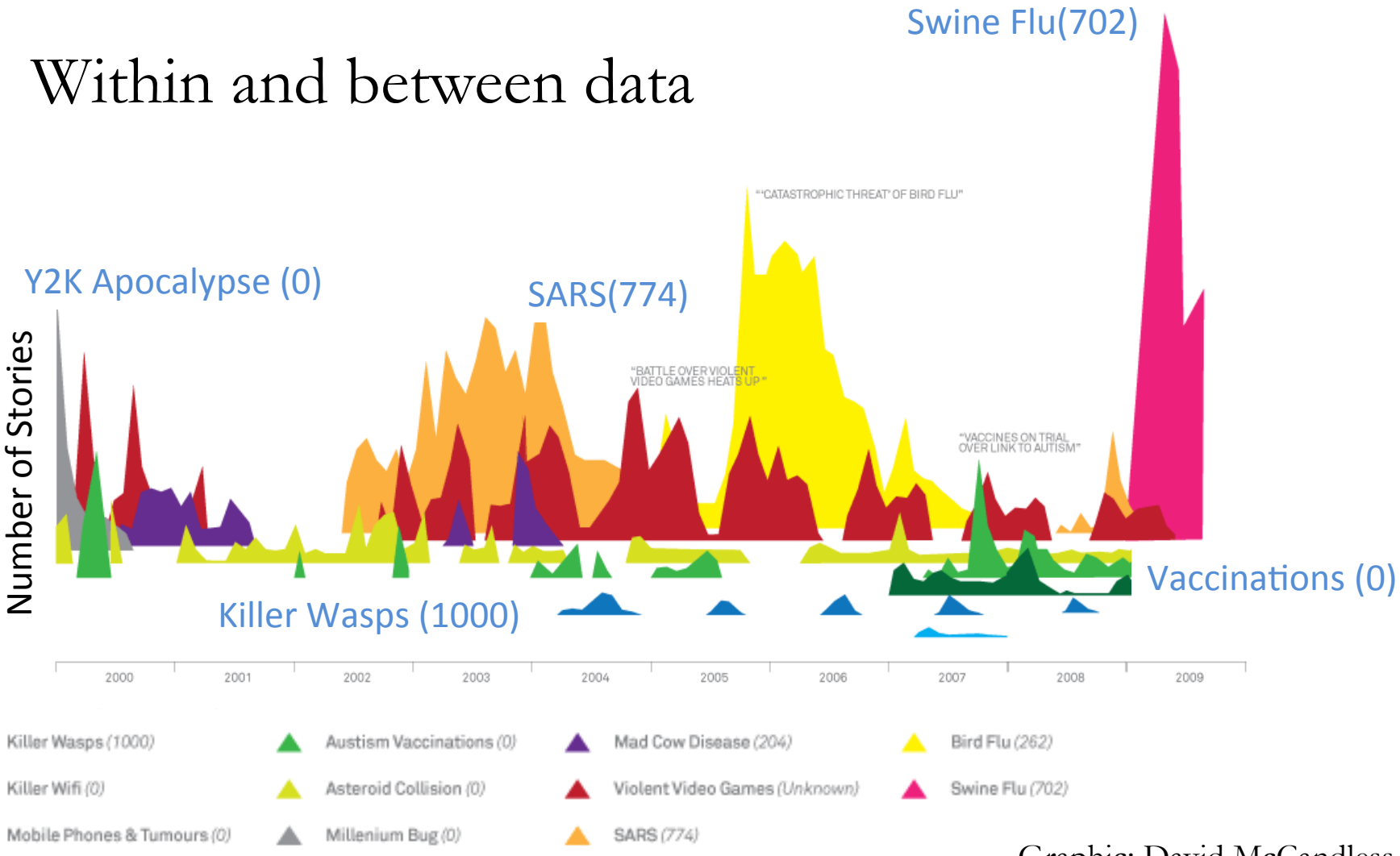
- Exposing Patterns



Graphic: David McCandless

Why Data Visualization?

- Within and between data



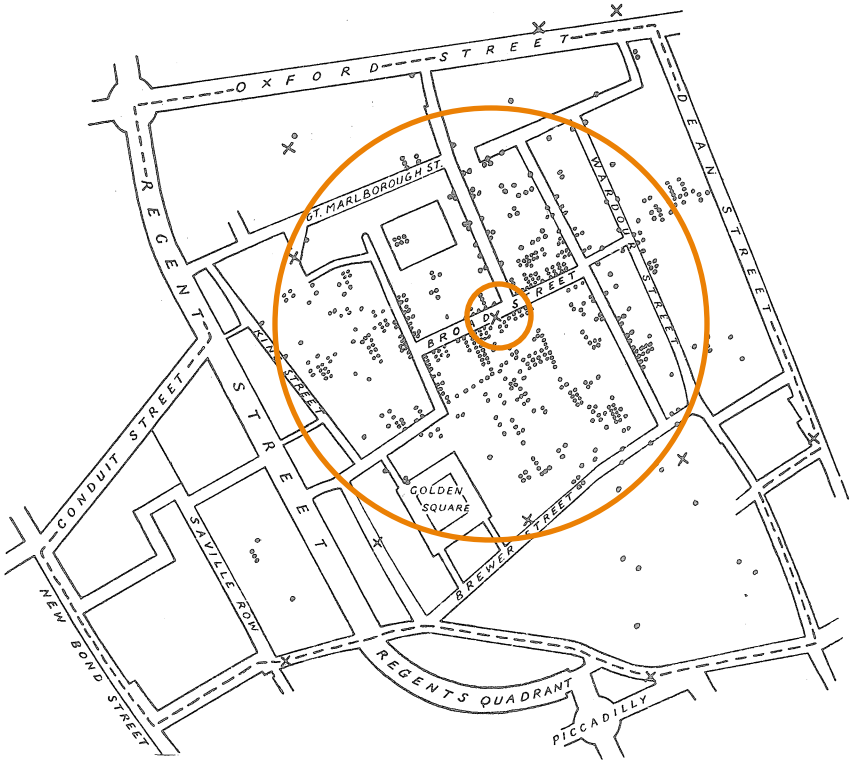
Graphic: David McCandless

Why Data Visualization?

- Visualizing data
  - Indicate issues or areas of further investigation
    - Important for all stages of analysis
  - Expose patterns and connections within and between data
  - Identify and show most important focus for a set of results

Why Data Visualization?

- Most important focus for set of results

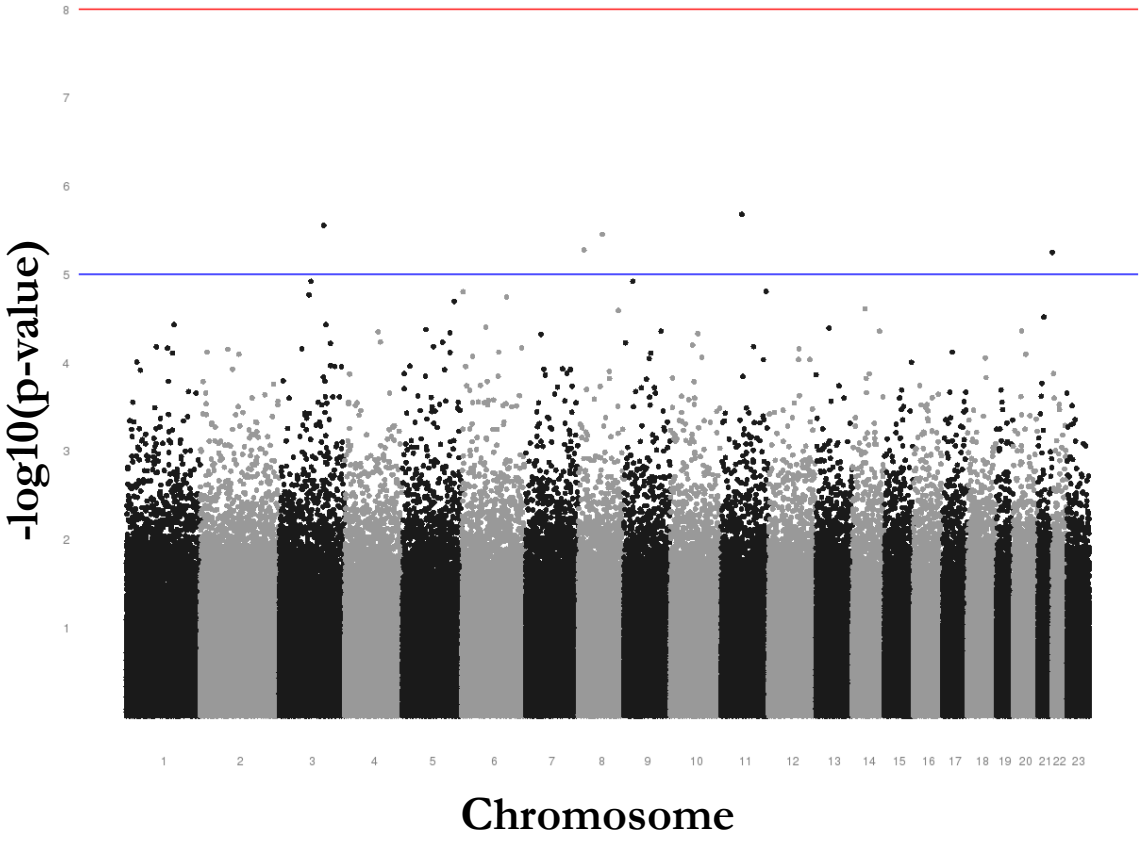


- John Snow and a map of a cholera outbreak - 1854
- Believed cholera transmission was not due to "bad air"
- Snow used this spot map to show others how cases of cholera were clustered around the Broad Street water pump

The result of the inquiry, then, is, that there has been no particular outbreak or prevalence of cholera in this part of London except among the persons who were in the habit of drinking the water of the above-mentioned pump well. – John Snow

Why Data Visualization?

- Most important focus for set of results



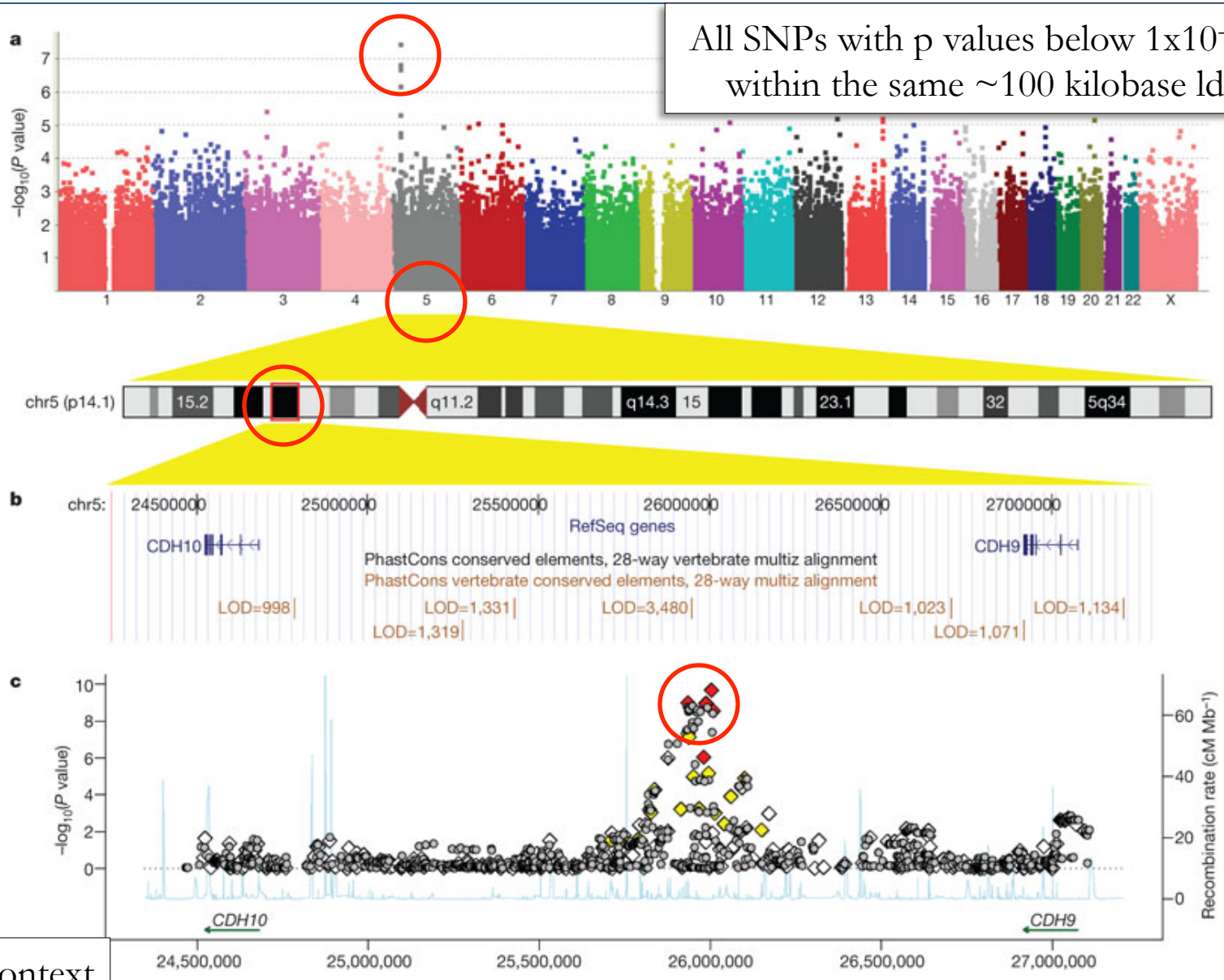
Graphic: http://gettinggeneticsdone.blogspot.com

Why Data Visualization?

- Example
  - Genome wide association studies for Autism Spectrum disorders – two separate cohorts of European ancestry
  - Six single nucleotide polymorphisms identified between genes *CDH10* and *CDH9*

Why Data Visualization?

All SNPs with p values below $1x10^{-7}$ resided within the same ~100 kilobase ld block

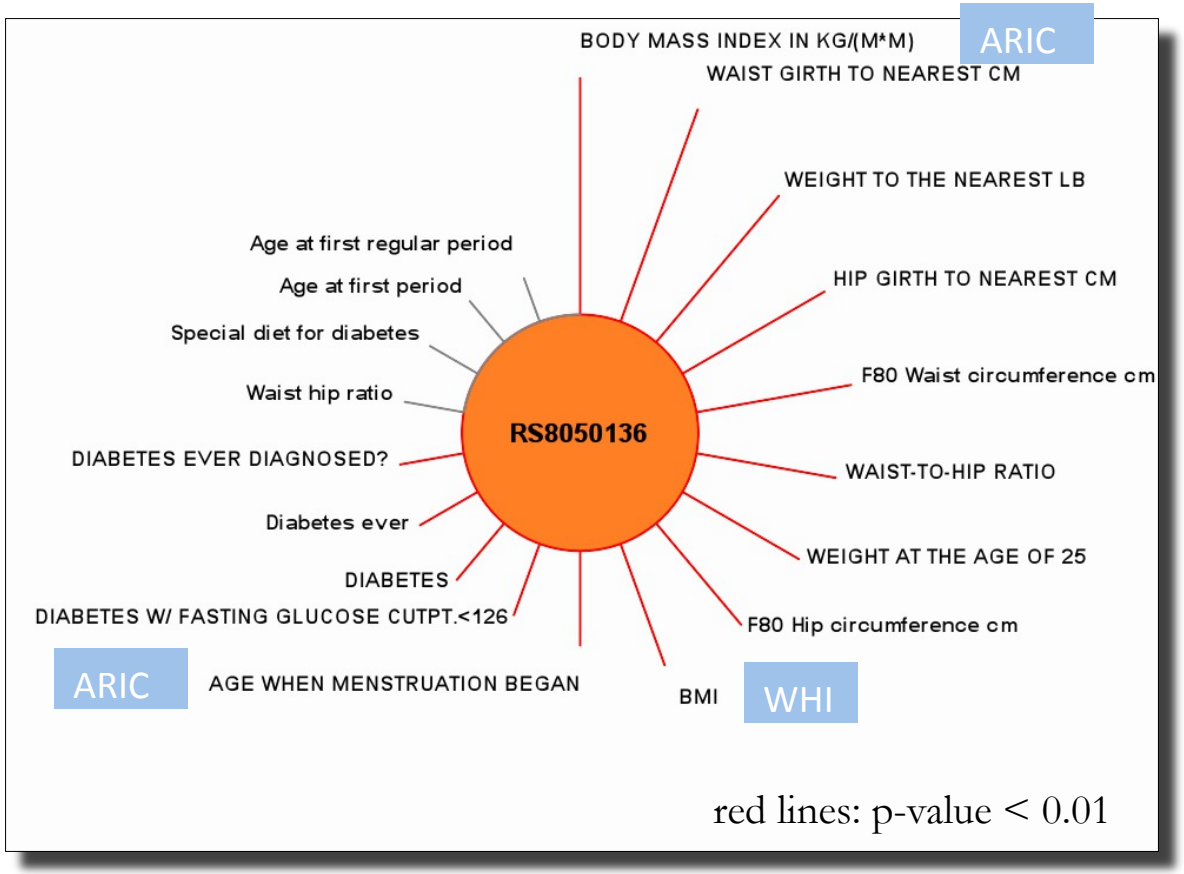


Spatial Context

Graphic: Wang et al. 2009

Why Data Visualization?

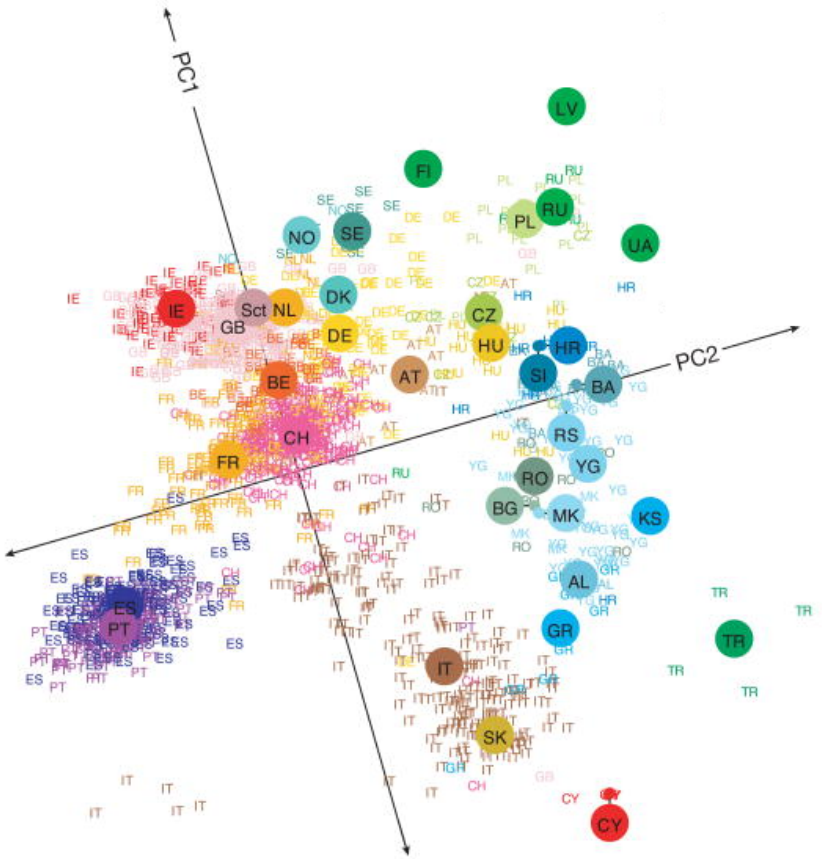
- Most important focus for set of results

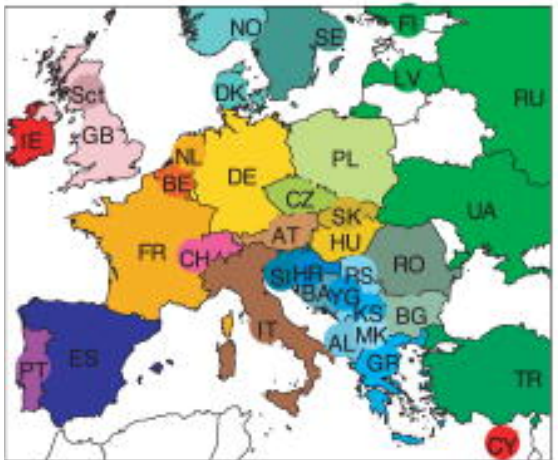


- Example: PheWAS
  - Phenome-wide association study
  - 84 SNPs, thousands of phenotypes, across studies
  - Result for one SNP across multiple phenotypes

Why Data Visualization?

- Identify most important set of results



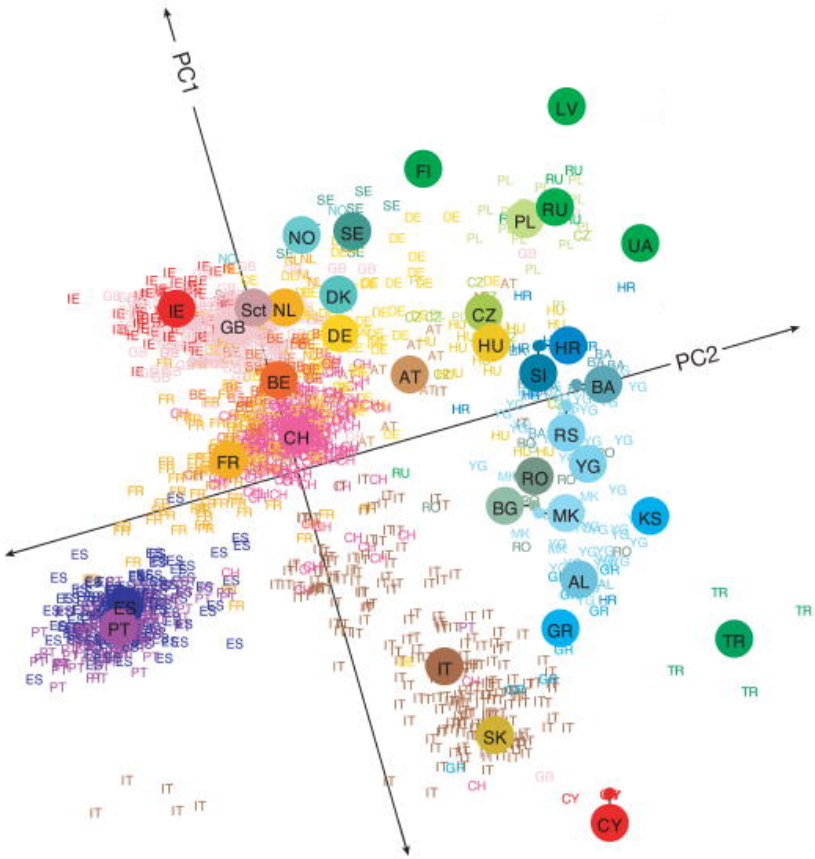


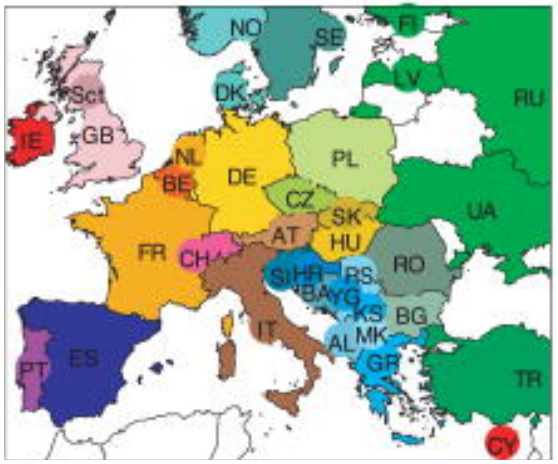
- 197,146 loci in 1,387 individuals from Europe
- Used principle components analysis to cluster the data
- Individuals from the same geographic region cluster together - major populations are distinguishable

Novembre et al. 2008

Why Data Visualization?

- Identify most important set of results



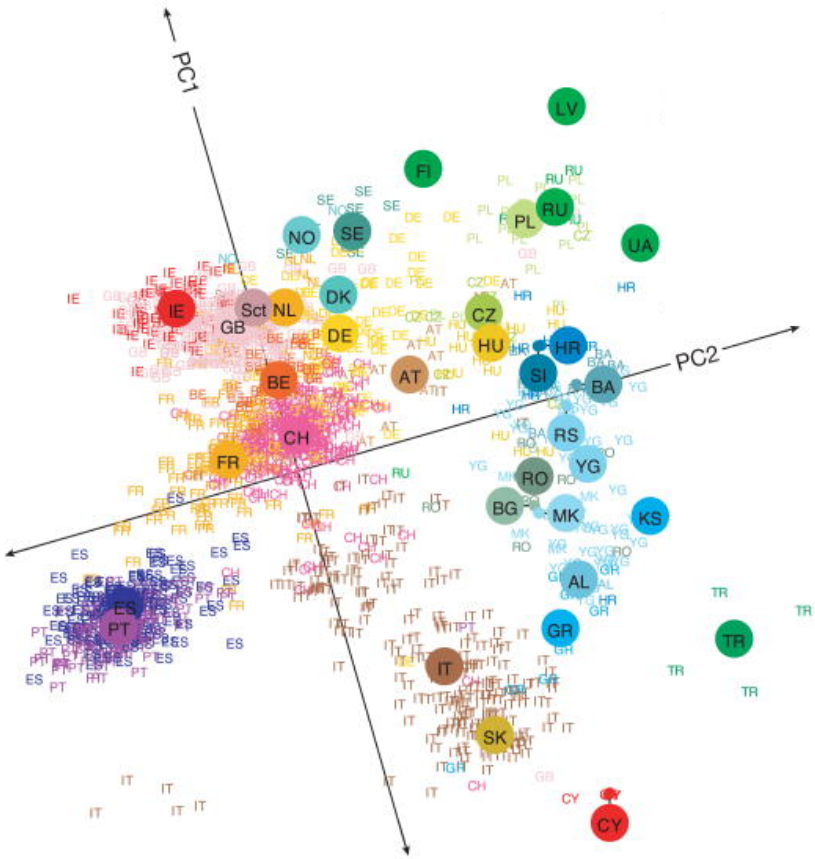


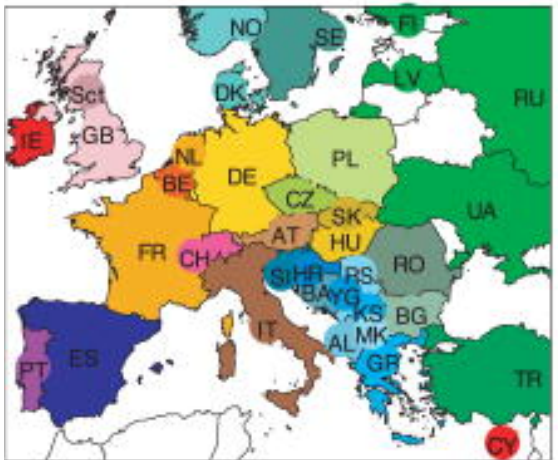
- The resulting two-dimensional plot resembles a geographic map of Europe

Novembre et al. 2008

Why Data Visualization?

- Identify most important set of results



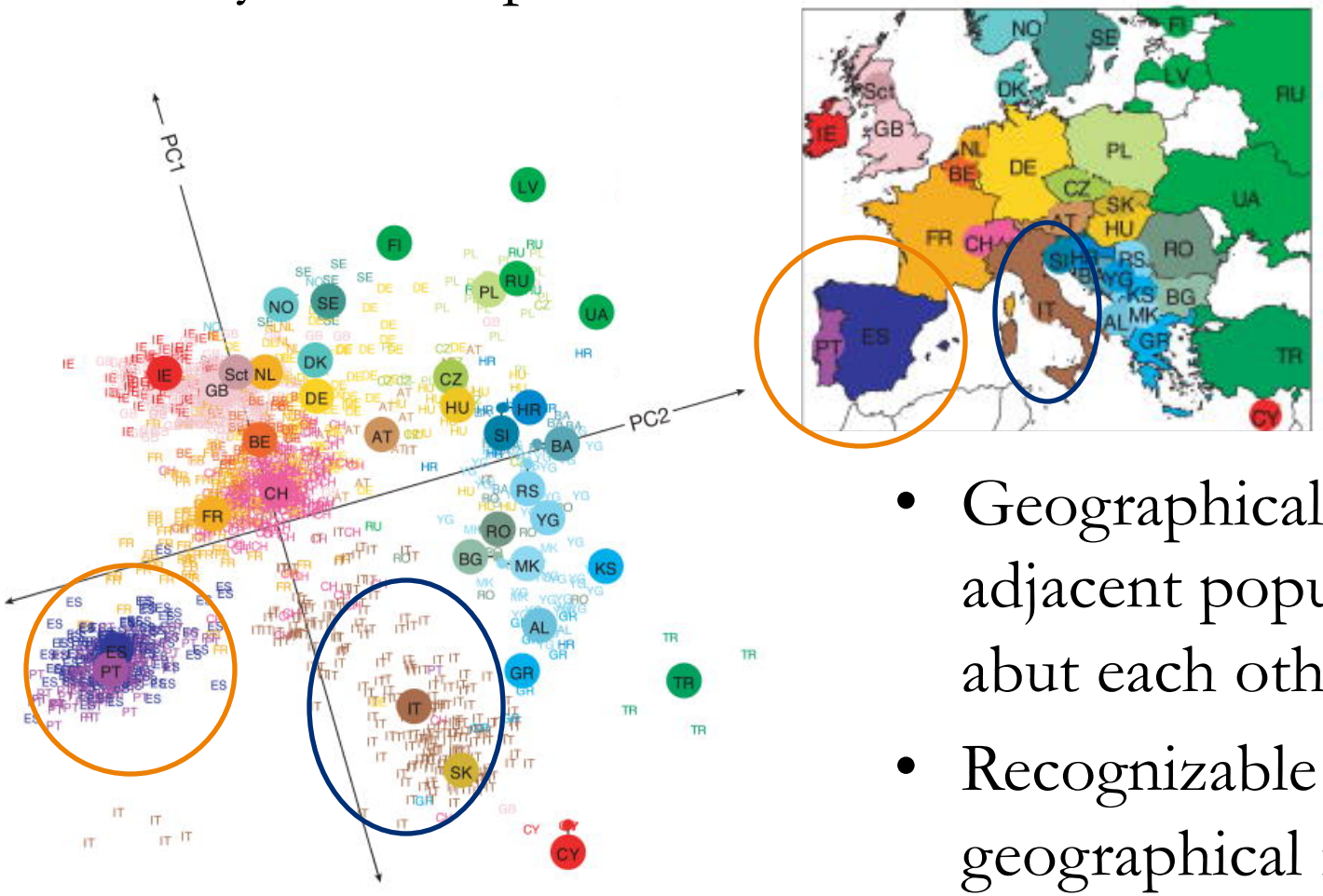


- The first two principle components accounting for the most variability in the data are effectively longitude/latitude

Novembre et al. 2008

Why Data Visualization?

- Identify most important set of results



- Geographically adjacent populations abut each other
- Recognizable geographical features

Novembre et al. 2008

Why Data Visualization?

- Visualizing data
  - Indicate issues or areas of further investigation
    - Important for all stages of analysis
  - Expose patterns and connections within and between data
  - Identify and show most important focus for a set of results
- Data can make more sense, expose trends, tell a story, or allow a specific focus

A Few Rules of Thumb

- Edward Tufte
  - Teaches courses in statistical evidence, information design, and interface design
  - A few rules of thumb
    - Choose whatever it takes to display info
    - "Don't get it original, get it right" in using conventional formats
    - Can anything be removed without erasing information?
    - Does your diagram/plot compare to a good map?
    - Evoke a content response from the start

A Few Rules of Thumb

- A few rules of thumb
  - WHAT is my QUESTION?
    - Let the question dictate your choice of plot
    - Don't let the software dictate your choice of plot
    - Axis? Temporal? Geographical?
  - After visualization ask yourself
    - Is there anything in this plot that is misleading?
  - Share, share, share your data visualization
    - Feedback!!!!
  - 7 Colors
    - What can you do with shapes, point size?

A Recommendation

- A few rules of thumb
  - WHAT is my QUESTION?
    - Let the question dictate your choice of plot
    - Don't let the software dictate your choice of plot

- How do I try out various types of data reduction or make it easier to ask questions about my data?

Databases

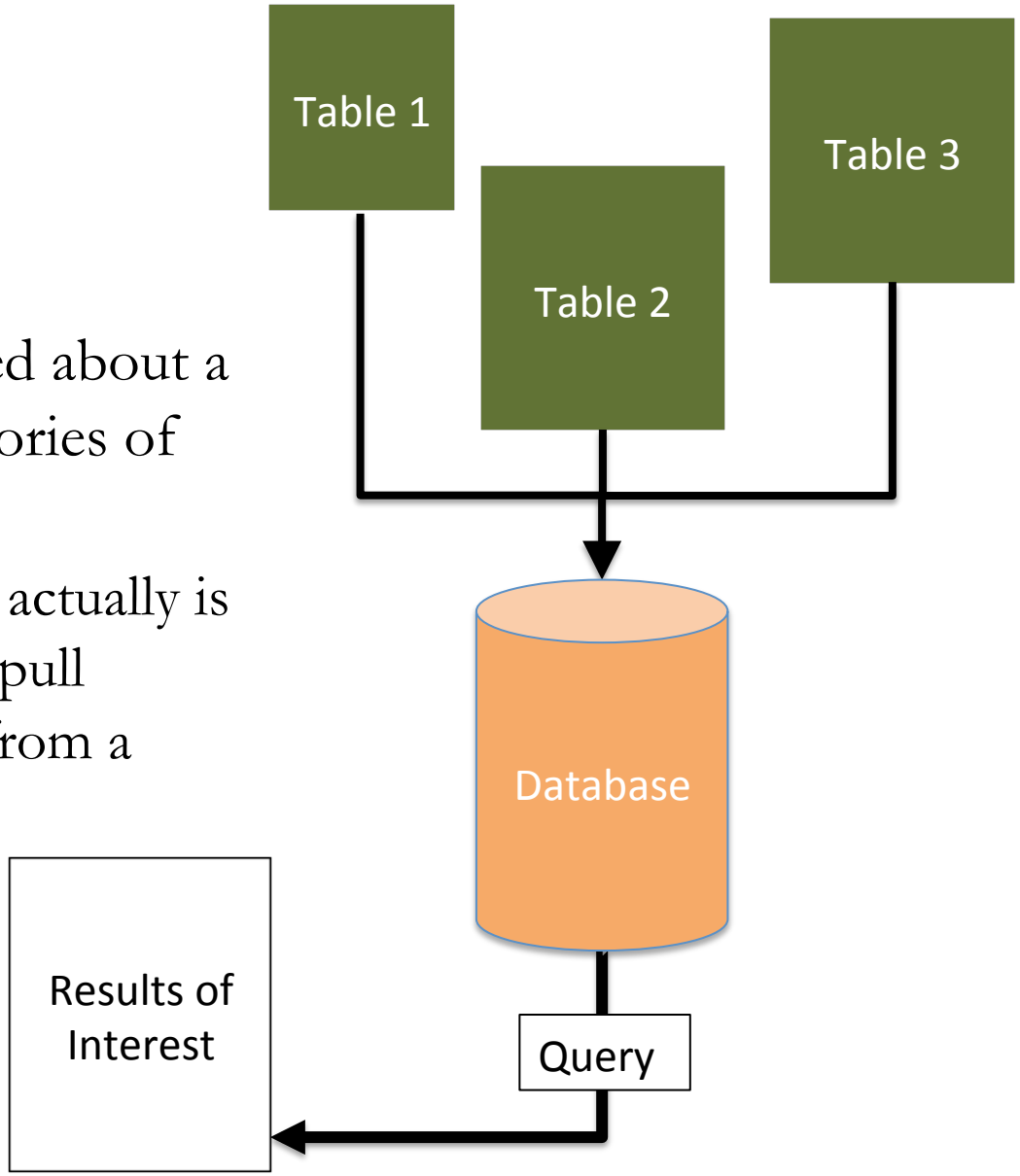
- We have all these software tools for organizing, analysis, and annotation of genomic data, what do we do with our results?

- How do I try out various types of data reduction or make it easier to ask questions about my data?
    - I could try to use Excel to spend some time perusing the data
        - Limited in what you can do
        - Whatever you do you have to do over and and over again
        - Hard to share
        - Can't scale up to extremely extensive data

- DATABASES!
    - Just like learning a little programming, learning how to create and use a database will be a very helpful tool

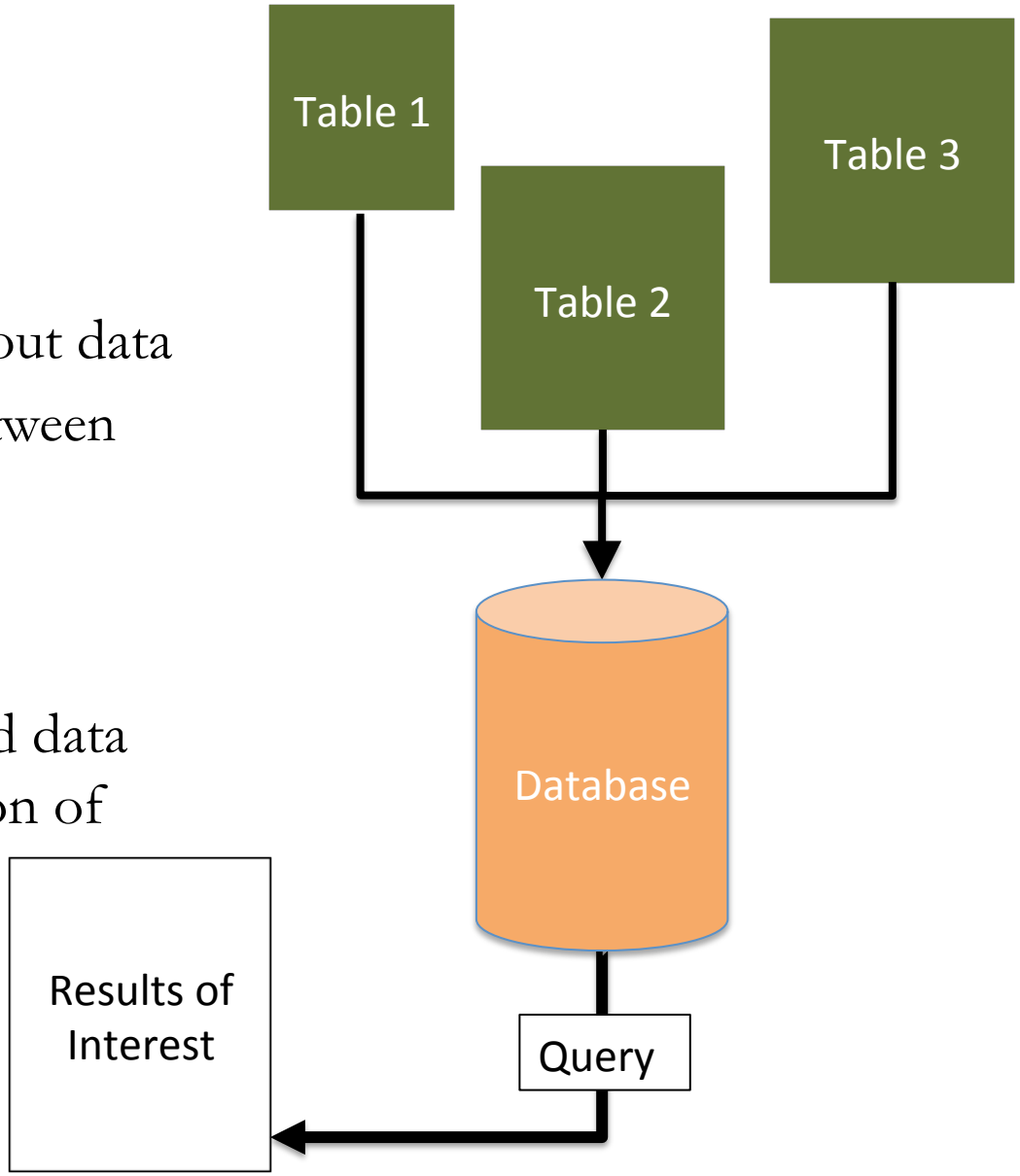
Databases

- So what is a database?
  - Obviously we have talked about a bunch of public repositories of data
    - The interface you use actually is generating a query to pull specific information from a database



Databases

- Relational databases
  - Store information about data
    - and connections between data
  - Tables with rows and columns
  - Way to retrieve related data based on your question of interest



Databases

- Once your data is your database, you can share that exact database with others

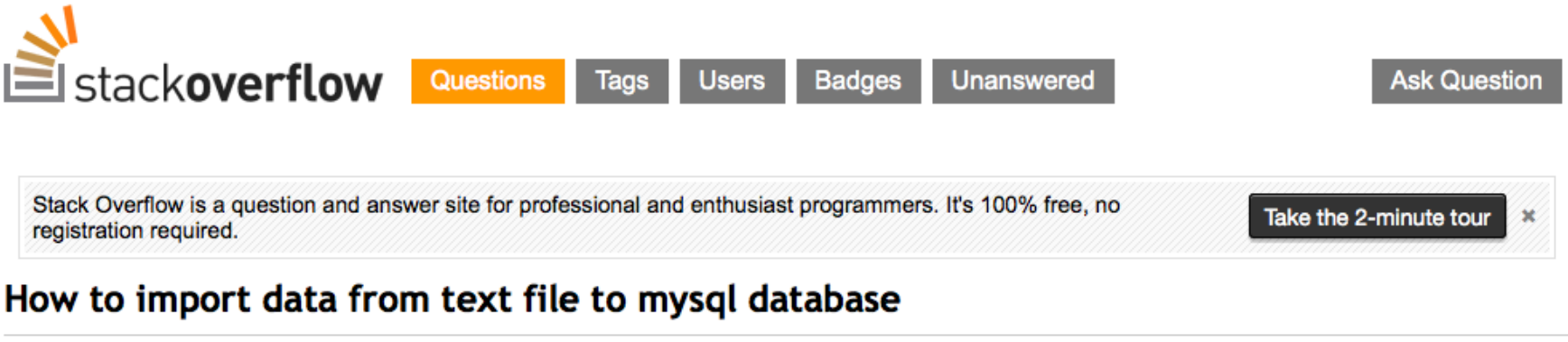
- Database queries
  - How you ask your question
  - Generally user friendly and interpretable
    - SELECT *information* FROM *database* WHERE *criteria*
    - Show me only results
      - p-value cutoff of interest
      - In a chromsome region (bp to bp)
      - With only specific type of annotation

Databases

- Multiple tables of information, using queries to bring together information filtered and interpretable
- Good idea to keep documentation about what you added and when

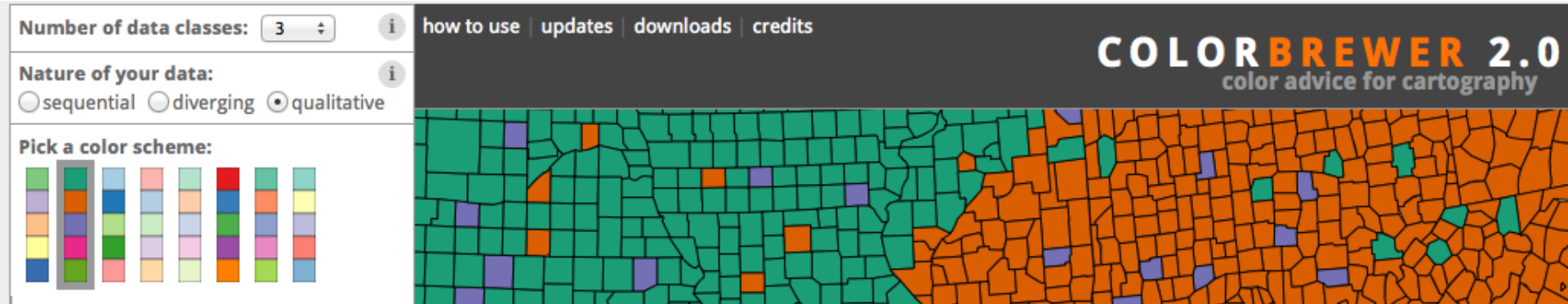
Databases

- How do I begin??
  - MySQL http://www.mysql.com/
    - http://dev.mysql.com/doc/refman/5.0/en/tutorial.html
  - Self-paced mini-course through Stanford
    - Free!
    - https://class.stanford.edu/courses/DB/2014/SelfPaced/about



Color Brewer

- 7 Colors
- Color Brewer: http://colorbrewer2.org/
- Color Brewer provides assistance for picking effective color combinations
  - Contrasts between colors
  - Avoiding red/green color combinations
  - Choosing effective grey scale tones

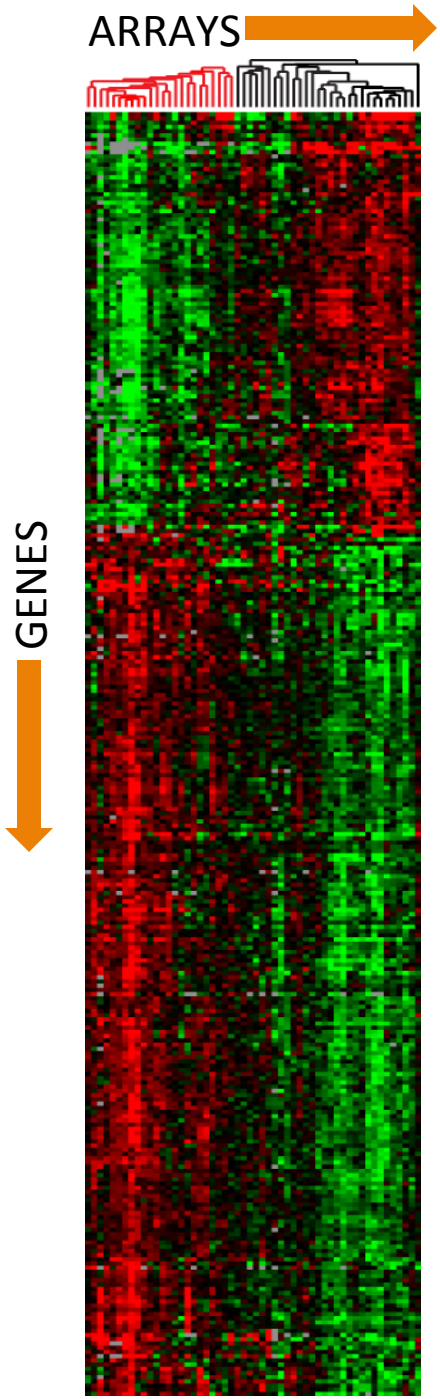


Breaking out of Tables

- What tables do
  - Show detail for all observations presented
  - Useful for curiosity about specific points
- What tables don't do
  - Limited to comparing only a certain amount of information within a table
  - "That is why they call it the big picture, rather than the big table"

Breaking out of Tables

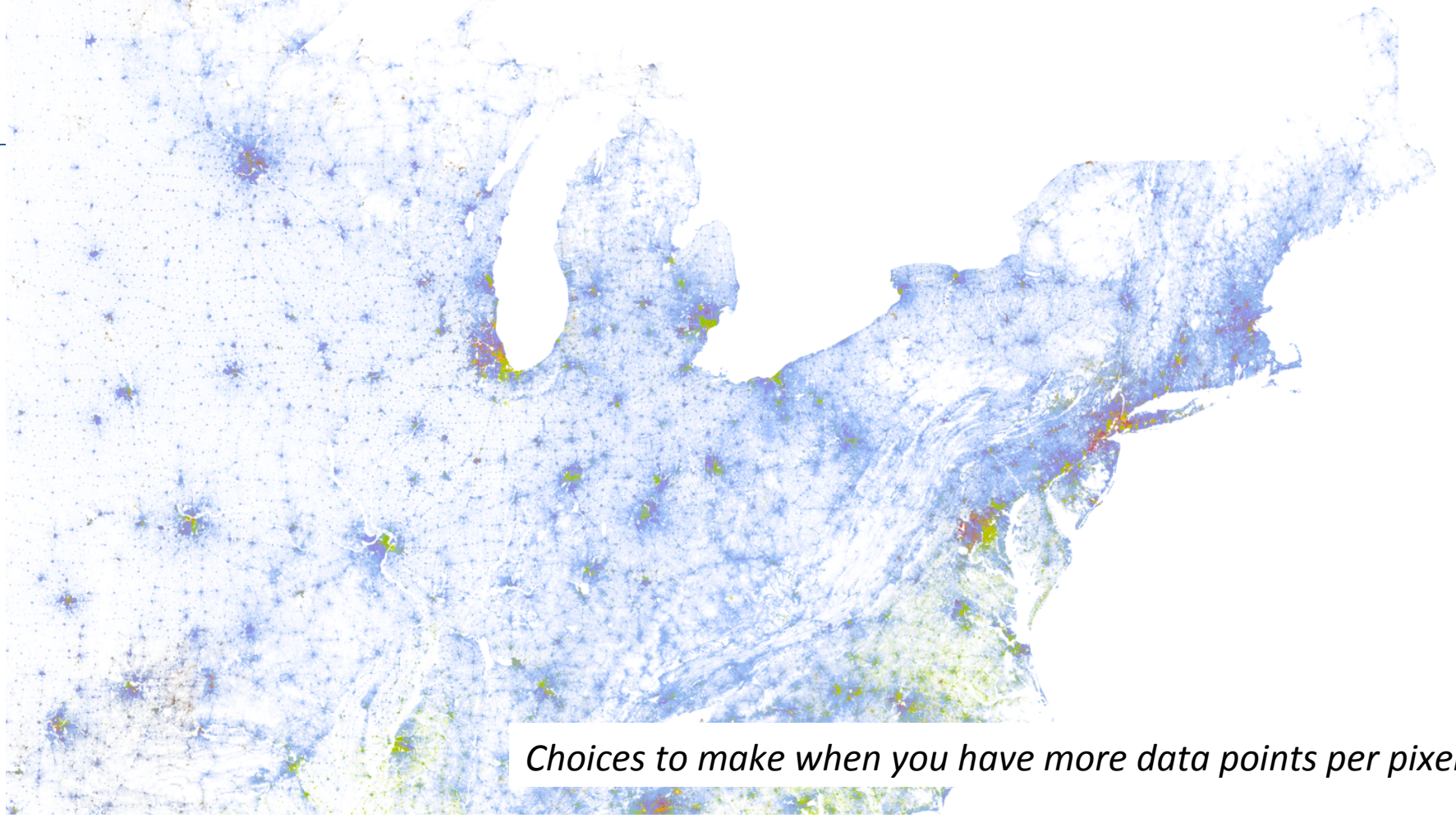
- Example: gene expression heat map
  - Microarray data (mRNA expression levels) could be presented in tables
  - Heat maps were devised to move to a more visual way of viewing the expression of thousands of genes



Particularly with huge amounts of data, don't try to show ALL the details

Ex. Dropping gene names on heat map

Graphic: Pendergrass et al. 2010



Choices to make when you have more data points per pixel

The Map

This map is an American snapshot; it provides an accessible visualization of geographic distribution, population density, and racial diversity of the American people in every neighborhood in the entire country. The map displays 308,745,538 dots, one for each person residing in the United States at the location they were counted during the 2010 Census. Each dot is color-coded by the individual's race and ethnicity. The map is presented in both black and white and full color versions. In the color version, each dot is color-coded by race.

http://www.coopercenter.org/demographics/Racial-Dot-Map

Breaking out of Tables

- Example: SNP-phenotype association results
  - Primarily presented in tabular form in both journals and presentations
  - Data visualization can make trends more evident in interpreting as well as sharing results

Breaking out of Tables

- Example: SNP-phenotype association results
  - BioVU proof-of-principle project
    - Ritchie et al. 2010
    - Question
      - Can EMR based data be used to replicate previously reported SNP/phenotype associations?
    - Genotyped SNPs in the first ~10,000 BioVU samples accrued

Breaking out of Tables

- Example: SNP-phenotype association results
  - BioVU proof-of-principle project
    - 21 SNPs

    - Algorithms used to define cases and controls in SD for 5 diseases
      - Atrial fibrilation
      - Crohn's disease
      - Multiple sclerosis
      - Rheumatoid arthritis
      - Type 2 diabetes

  - Content experts guided algorithm development
    - ICD-9 billing codes, medication usage information, free-text

Breaking out of Tables

- Example: SNP-phenotype association results
  - BioVU proof-of-principle project
    - Multiple ways of defining cases
      - Atrial fibrillation, Crohn's disease, type 2 diabetes
        - EA with definite cases
        - EA + unknown ancestry with definite cases
      - Rheumatoid arthritis, multiple sclerosis
        - EA with definite cases
        - EA with definite and probable cases
          - Probable: potential overlapping with other disease
        - EA + unknown with definite cases
        - EA + unknown with definite and probable cases

Breaking out of Tables

Table 1. Allelic Odds Ratio

SNP	Cases N	Controls N	Cases Minor Allele Frequency	Cases Minor Allele	Controls Minor Allele Frequency	Controls Minor Allele	Controls Allelic Chi-Square p Value	Odds Ratio	95% Confidence Interval
ATRIAL FIBRILLATION									
Ancestry: Non-Hispanic European; Case Definition: Definite									
rs2200733	147	1439	0.1599	T	0.1120	T	0.0147	1.5093	(1.08–2.11)
rs10033464	143	1402	0.1084	T	0.0881	T	0.2530	1.2585	(0.85–1.87)
Ancestry: Non-Hispanic European + Unknown; Case Definition: Definite									
rs2200733	148	1467	0.1588	T	0.1115	T	0.0153	1.5048	(1.08–2.10)
rs10033464	144	1432	0.1111	T	0.0887	T	0.2066	1.2844	(0.87–1.90)
CROHN DISEASE									
Ancestry: Non-Hispanic European; Case Definition: Definite									
rs11805303	107	2884	0.3271	T	0.3017	T	0.4263	1.1253	(0.84–1.51)
rs17234657	106	2890	0.2028	G	0.1201	G	0.0003	1.8646	(1.32–2.63)
rs1000113	107	2905	0.0935	T	0.0730	T	0.2601	1.3096	(0.82–2.10)
rs17221417	107	2896	0.3785	G	0.2949	G	0.0086	1.4562	(1.10–1.93)
rs2542151	107	2901	0.1542	G	0.1649	G	0.6774	1.0834	(0.74–1.58)
Ancestry: Non-Hispanic European + Unknown; Case Definition: Definite									
rs11805303	110	3175	0.3288	T	0.3054	T	0.4557	1.1145	(0.84–1.45)
rs17234657	110	3182	0.1955	G	0.1204	G	0.0009	1.7756	(1.23–2.50)
rs1000113	111	3199	0.0946	T	0.0769	T	0.3323	1.2542	(0.79–1.98)
rs17221417	111	3188	0.3874	G	0.2917	G	0.0021	1.5353	(1.17–2.02)
rs2542151	111	3195	0.1577	G	0.1635	G	0.8158	1.0446	(0.72–1.51)
MULTIPLE SCLEROSIS									
Ancestry: Non-Hispanic European; Case Definition: Definite									
rs6897932	61	1861	0.2049	T	0.2515	T	0.2425	1.3036	(0.83–2.04)
rs3135388	61	1892	0.2887	T	0.1427	T	<0.0001	2.3210	(1.55–3.48)
rs2104286	61	1888	0.2377	A	0.2582	A	0.6102	1.1163	(0.73–1.70)
Ancestry: Non-Hispanic European + Unknown; Case Definition: Definite									
rs6897932	88	2105	0.2045	T	0.2485	T	0.1855	1.2857	(0.89–1.87)
rs3135388	88	2139	0.2955	T	0.1431	T	<0.0001	2.5120	(1.80–3.51)
rs2104286	88	2133	0.2102	A	0.2586	A	0.1503	1.3101	(0.91–1.89)
Ancestry: Non-Hispanic European; Case Definition: Definite + Probable									
rs6897932	68	1861	0.2132	T	0.2515	T	0.3118	1.2396	(0.82–1.88)
rs3135388	68	1892	0.2574	T	0.1427	T	0.0002	2.0818	(1.40–3.09)
rs2104286	68	1888	0.2279	A	0.2582	A	0.4275	1.1790	(0.78–1.77)
Ancestry: Non-Hispanic European + Unknown; Case Definition: Definite + Probable									
rs6897932	96	2105	0.2083	T	0.2485	T	0.2072	1.2563	(0.88–1.79)
rs3135388	96	2139	0.2760	T	0.1431	T	<0.0001	2.2840	(1.65–3.17)
rs2104286	96	2133	0.2083	A	0.2586	A	0.1190	1.3252	(0.93–1.89)

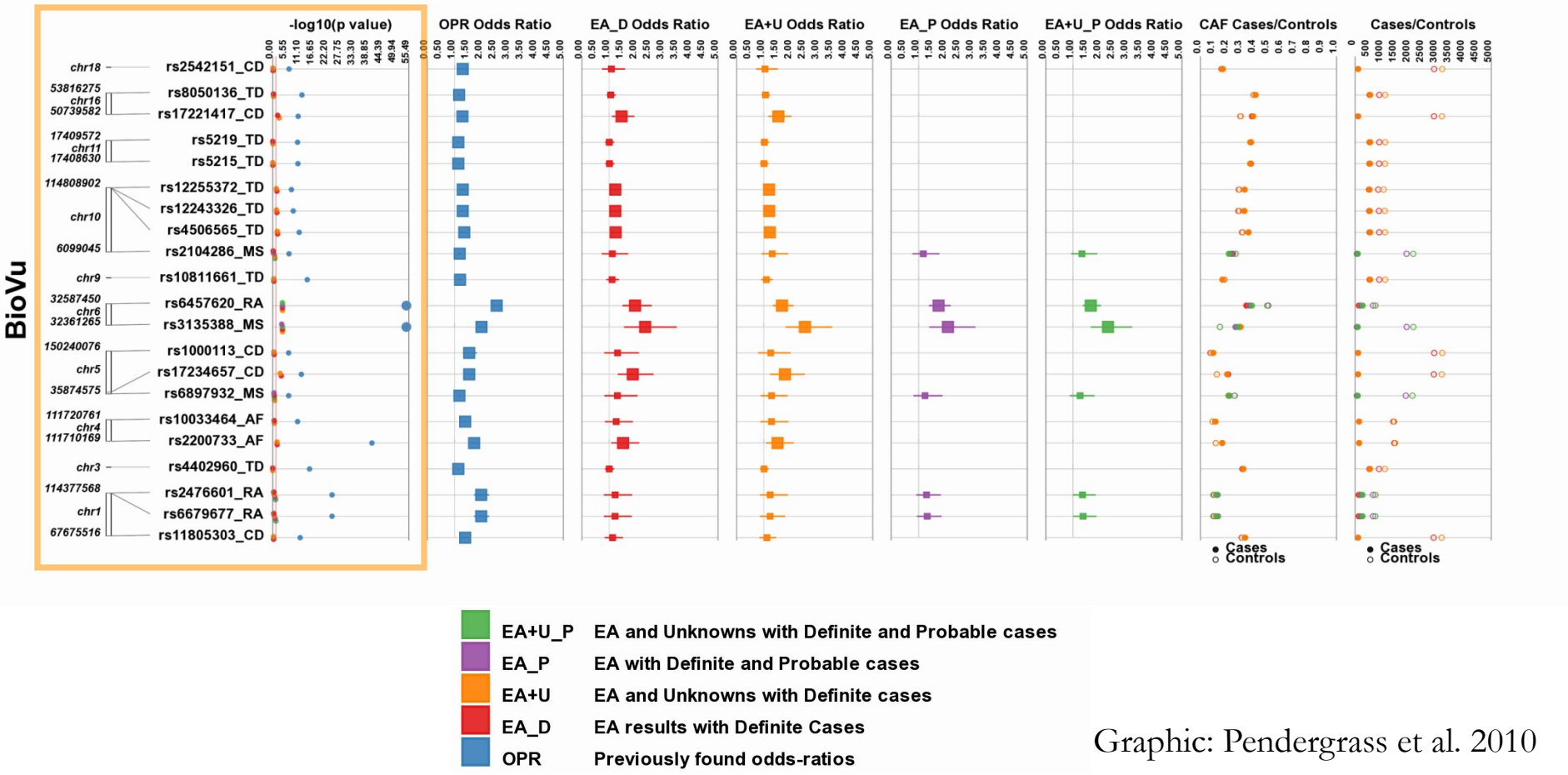
Table 1. *Continued*

SNP	Cases N	Controls N	Cases Minor Allele Frequency	Cases Minor Allele	Controls Minor Allele Frequency	Controls Minor Allele	Controls Allelic Chi-Square p Value	Odds Ratio	95% Confidence Interval
RHEUMATOID ARTHRITIS									
Ancestry: Non-Hispanic European; Case Definition: Definite									
rs6679677	134	658	0.1194	A	0.1003	A	0.3496	1.2162	(0.81–1.83)
rs2476601	134	659	0.1194	A	0.1002	A	0.3454	1.2183	(0.81–1.84)
rs6457620	138	662	0.3370	T	0.4977	T	<0.0001	1.9501	(1.49–2.56)
Ancestry: Non-Hispanic European + Unknown; Case Definition: Definite									
rs6679677	184	745	0.1141	A	0.0943	A	0.2609	1.2326	(0.86–1.78)
rs2476601	184	746	0.1141	A	0.0945	A	0.2576	1.2344	(0.86–1.88)
rs6457620	188	750	0.3601	T	0.4973	T	<0.0001	1.6689	(1.33–2.09)
Ancestry: Non-Hispanic European; Case Definition: Definite + Probable									
rs6679677	210	658	0.1286	A	0.1003	A	0.1029	1.3235	(0.94–1.85)
rs2476601	210	659	0.1262	A	0.1002	A	0.1319	1.2975	(0.92–1.82)
rs6457620	214	662	0.3626	T	0.4977	T	<0.0001	1.7422	(1.39–2.18)
Ancestry: Non-Hispanic European + Unknown; Case Definition: Definite + Probable									
rs6679677	272	745	0.1250	A	0.0946	A	0.0459	1.3667	(1.00–1.86)
rs2476601	272	746	0.1232	A	0.0945	A	0.0589	1.3459	(0.99–1.83)
rs6457620	277	750	0.3776	T	0.4896	T	<0.0001	1.6521	(1.35–2.02)
TYPE 2 DIABETES									
Ancestry: Non-Hispanic European; Case Definition: Definite									
rs4402960	527	877	0.3083	T	0.3079	T	0.9787	1.0023	(0.85–1.18)
rs10811661	534	887	0.1610	C	0.1753	C	0.3269	1.1074	(0.90–1.36)
rs4506565	532	886	0.3524	T	0.3053	T	0.0093	1.2384	(1.05–1.46)
rs12243326	520	876	0.3212	C	0.2785	C	0.0169	1.2253	(1.04–1.45)
rs12255372	510	847	0.3245	T	0.2816	T	0.0178	1.2257	(1.04–1.45)
rs5215	527	882	0.3672	C	0.3702	C	0.8728	1.0130	(0.86–1.19)
rs5219	533	888	0.3715	T	0.3705	T	0.9580	1.0042	(0.86–1.18)
rs8050136	533	886	0.4053	A	0.3916	A	0.4731	1.0584	(0.91–1.24)
Ancestry: Non-Hispanic European + Unknown; Case Definition: Definite									
rs4402960	548	1089	0.3139	T	0.3159	T	0.9067	1.0094	(0.86–1.18)
rs10811661	555	1103	0.1604	C	0.1727	C	0.3700	1.0931	(0.90–1.33)
rs4506565	553	1100	0.3535	T	0.3100	T	0.0117	1.2172	(1.04–1.42)
rs12243326	541	1088	0.3226	C	0.2845	C	0.0251	1.1976	(1.02–1.40)
rs12255372	530	1048	0.3236	T	0.2863	T	0.0305	1.1928	(1.02–1.40)
rs5215	547	1098	0.3656	C	0.3643	C	0.9404	1.0057	(0.87–1.17)
rs5219	554	1103	0.3700	T	0.3649	T	0.7728	1.0223	(0.88–1.19)
rs8050136	554	1102	0.4043	A	0.3897	A	0.4177	1.0628	(0.92–1.23)

Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record, *Ritchie et al. 2010*

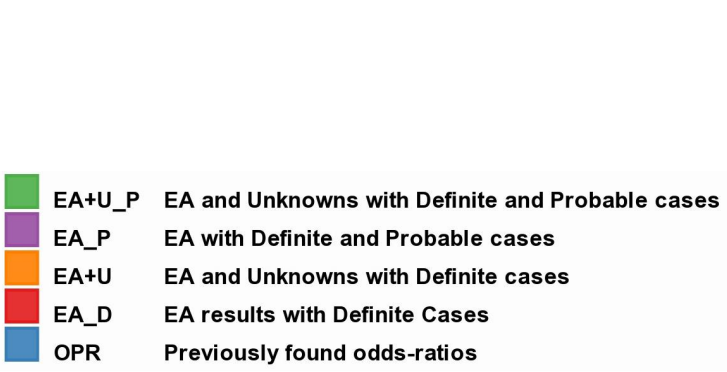
Breaking out of Tables

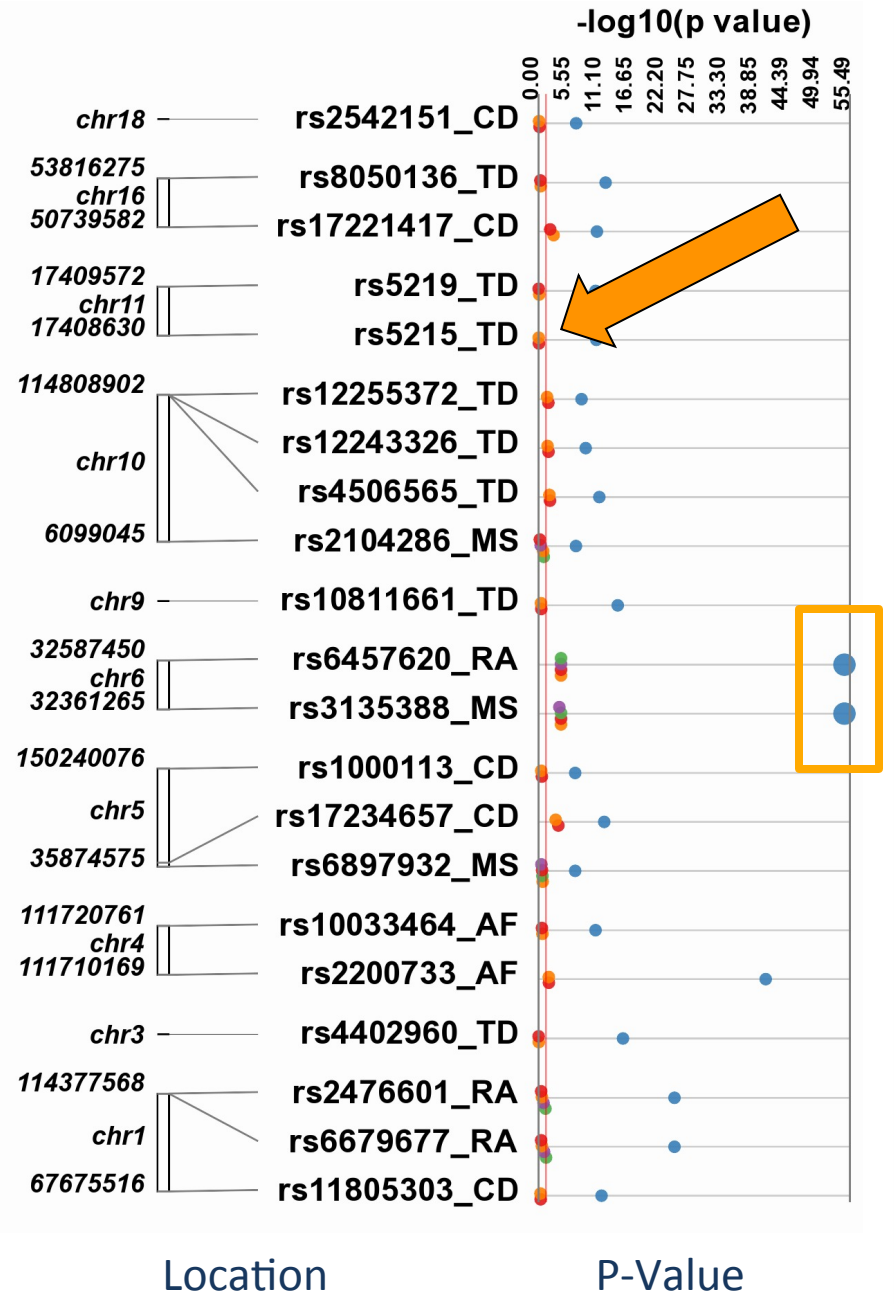
- Forest plot, with multiple tracks of data



Graphic: Pendergrass et al. 2010

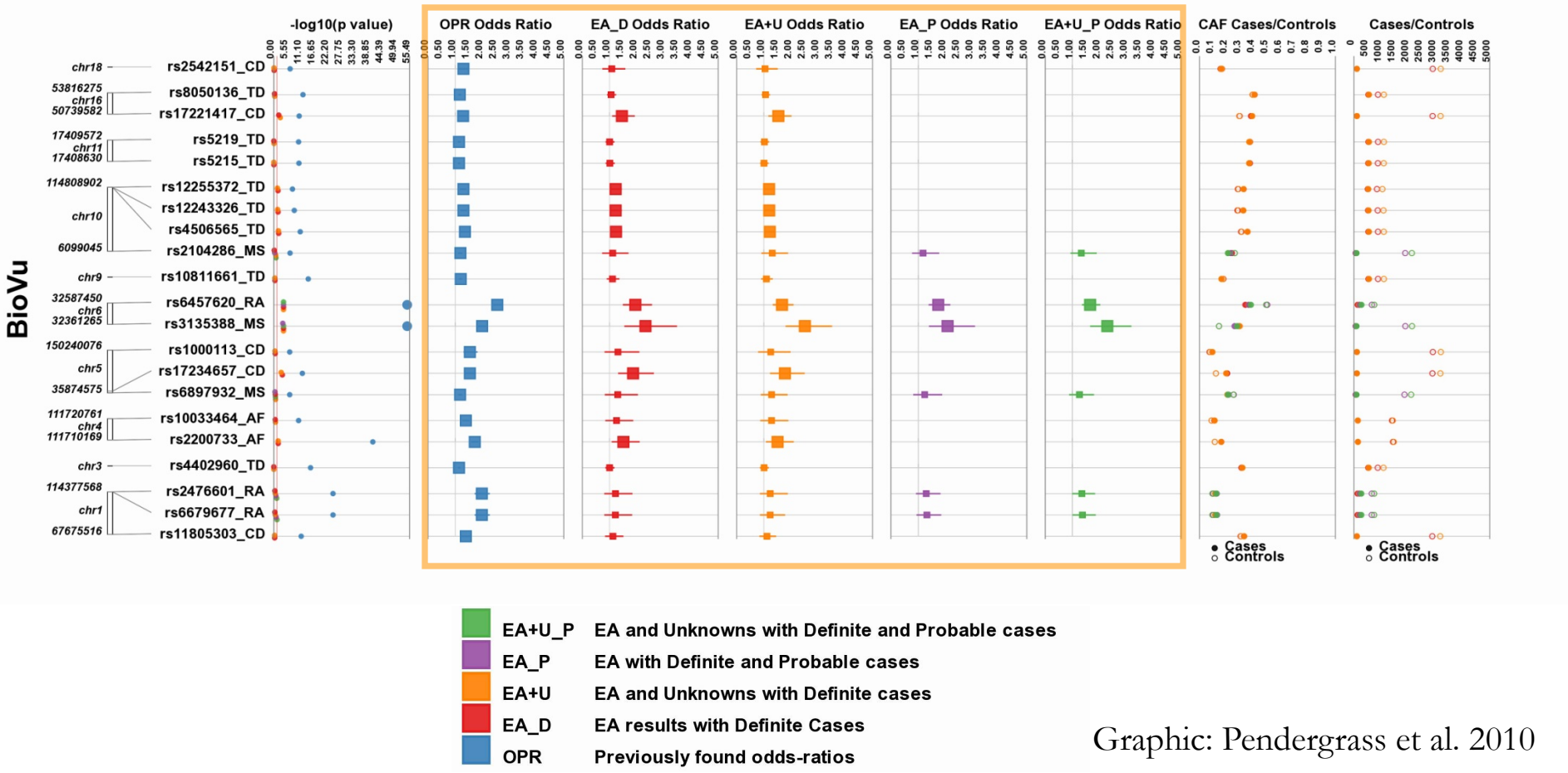
AF = atrial fibrillation
CD = Crohn's disease
RA = rheumatoid arthritis
TD = type 2 diabetes
MS = multiple sclerosis



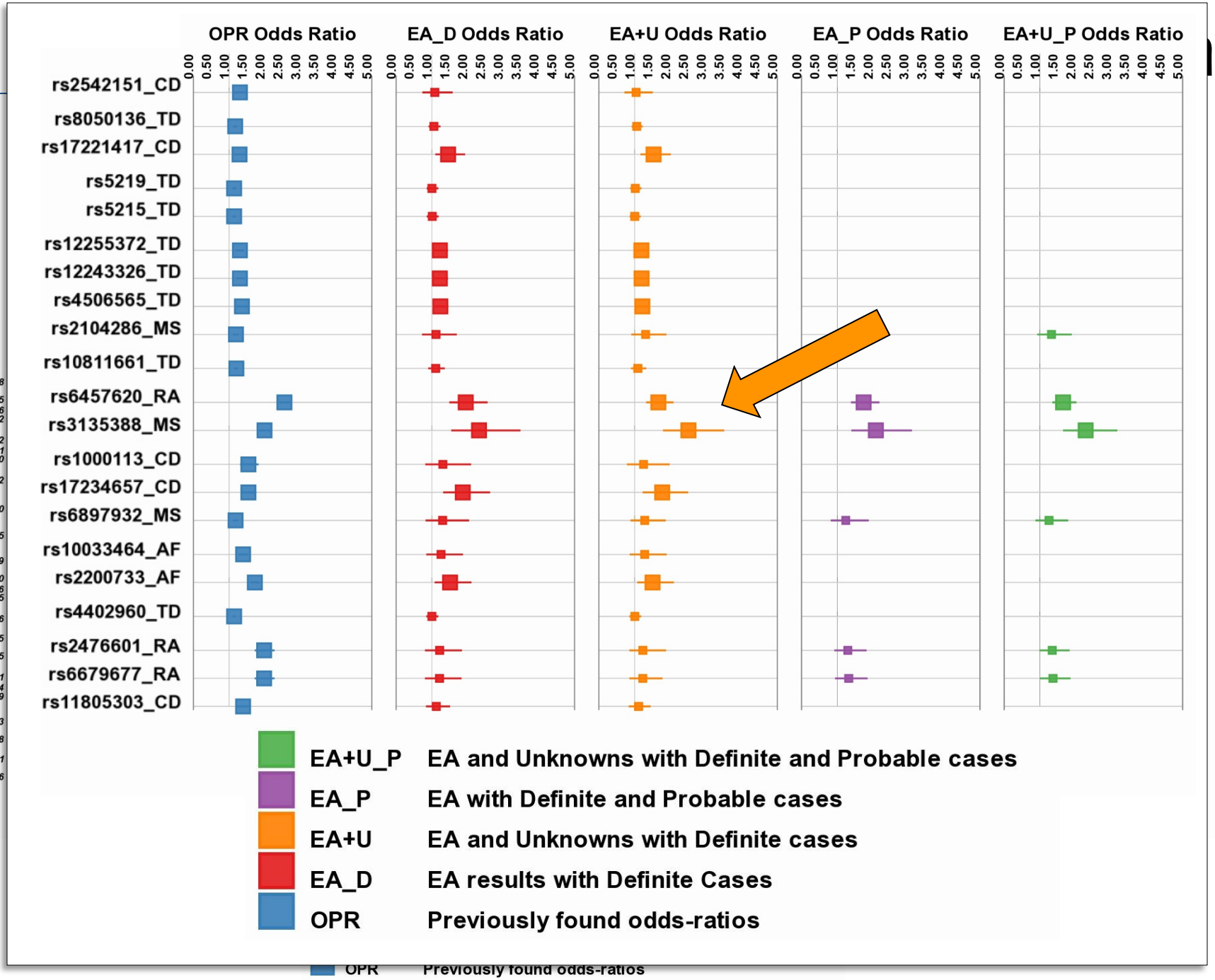


Breaking out of Tables

- Forest plot, with multiple tracks of data

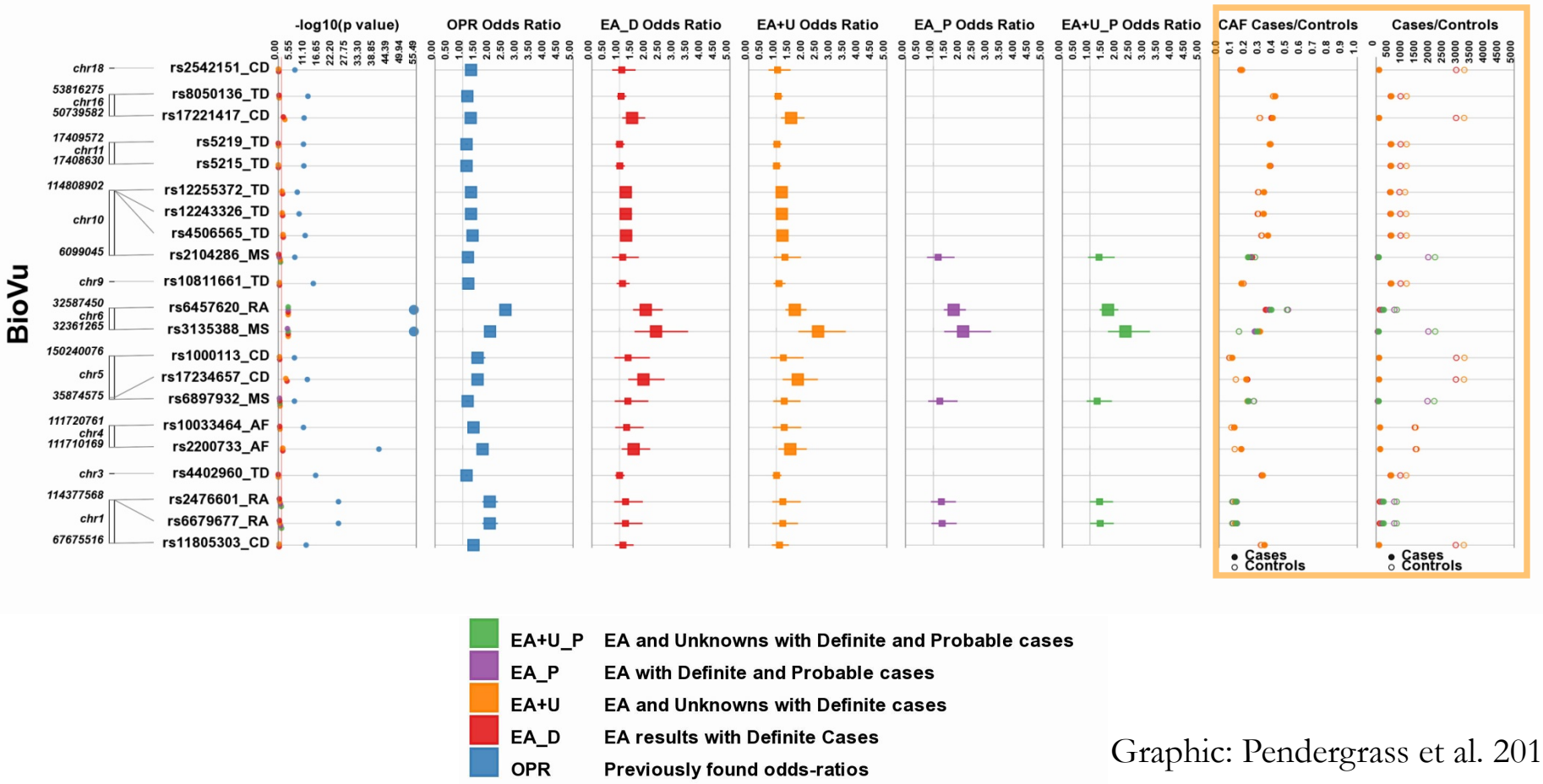


Graphic: Pendergrass et al. 2010

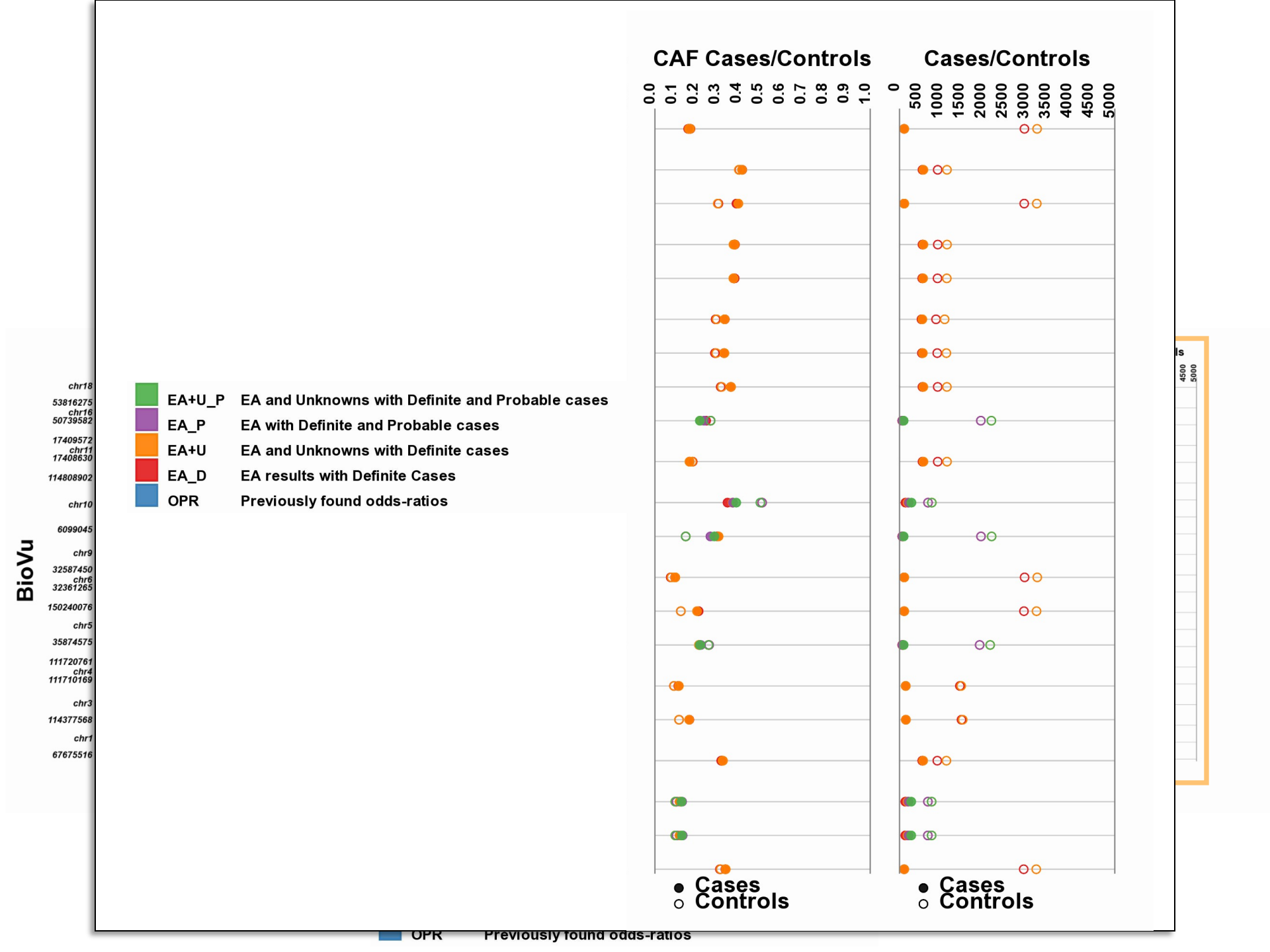

OPR Odds Ratio
EA_D Odds Ratio
EA+U Odds Ratio
EA_P Odds Ratio
EA+U_P Odds Ratio
0.00 0.50 1.00 1.50 2.00 2.50 3.00 3.50 4.00 4.50 5.00
rs2542151_CD
rs8050136_TD
rs17221417_CD
rs5219_TD
rs5215_TD
rs12255372_TD
rs12243326_TD
rs4506565_TD
rs2104286_MS
rs10811661_TD
chr18
53816275
rs6457620_RA
chr16
50739582
rs3135388_MS
17409572
chr11
17408630
rs1000113_CD
114808902
rs17234657_CD
chr10
rs6897932_MS
6099045
rs10033464_AF
chr9
32587450
rs2200733_AF
chr6
32361265
rs4402960_TD
150240076
chr5
rs2476601_RA
35874575
111720761
rs6679677_RA
chr4
111710169
rs11805303_CD
chr3
114377568
chr1
67675516
BioVu
EA+U_P EA and Unknowns with Definite and Probable cases
EA_P EA with Definite and Probable cases
EA+U EA and Unknowns with Definite cases
EA_D EA results with Definite Cases
OPR Previously found odds-ratios

Breaking out of Tables

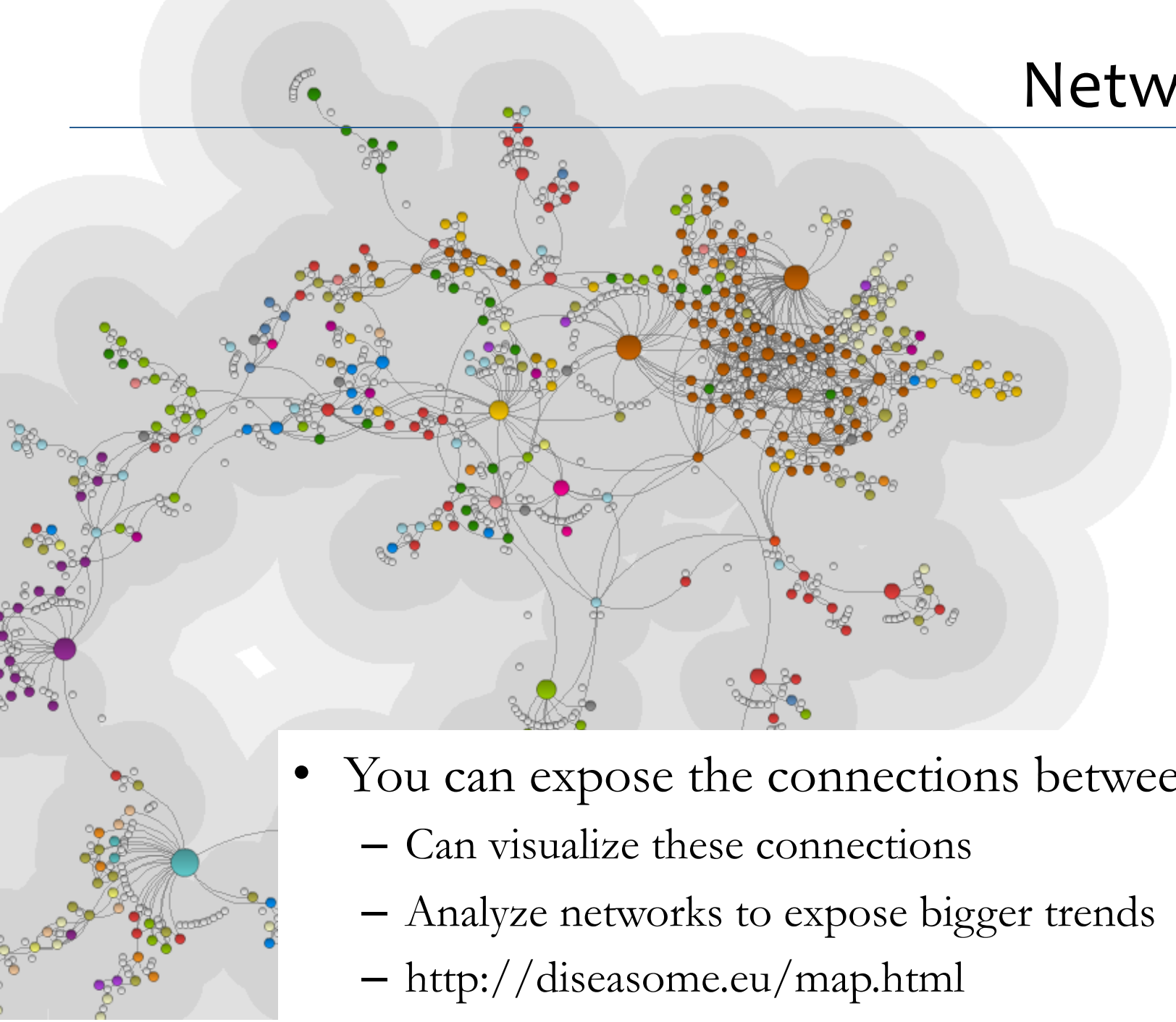
- Forest plot, with multiple tracks of data



Graphic: Pendergrass et al. 2010

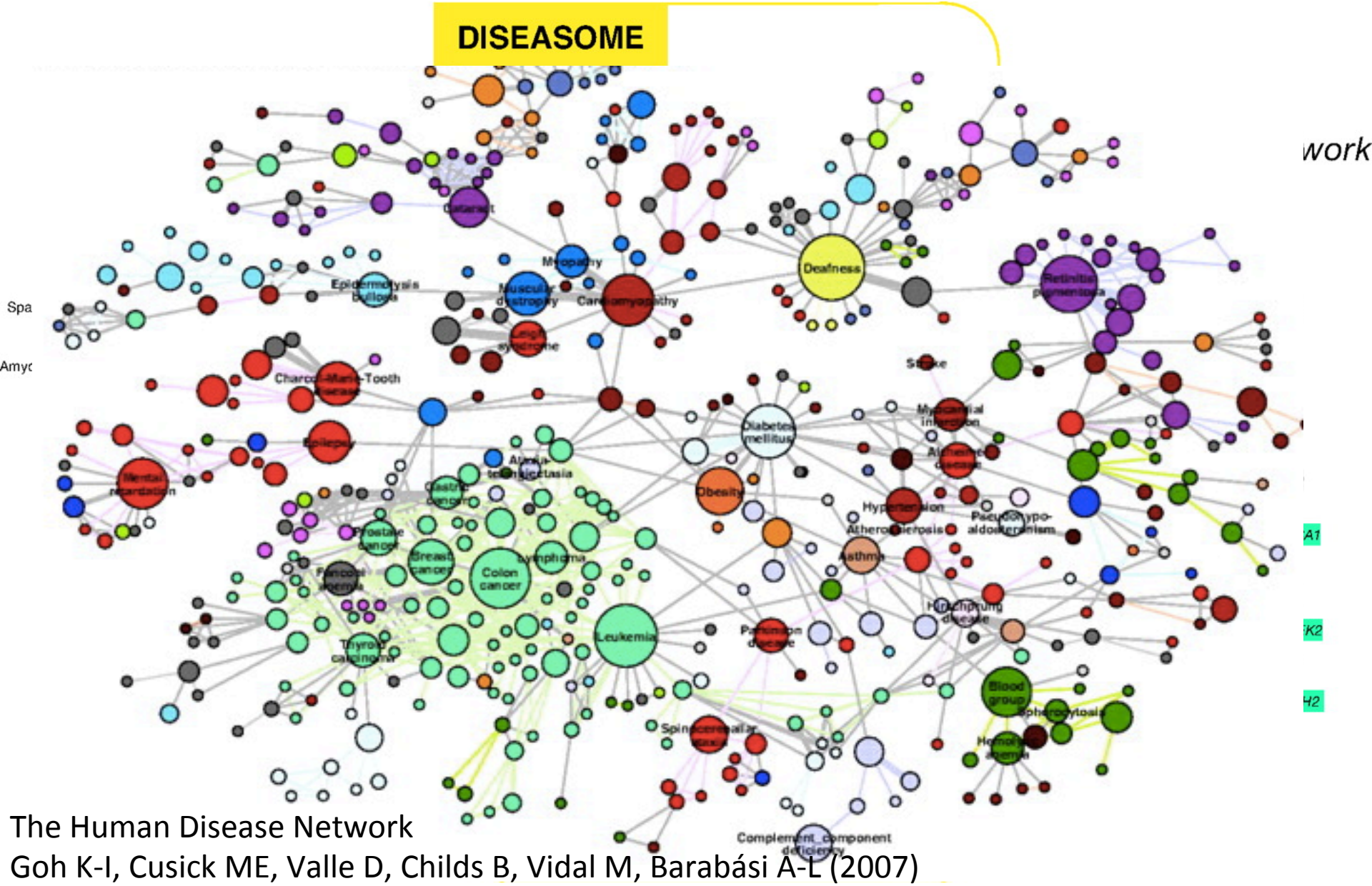

CAF Cases/Controls
Cases/Controls
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
0 500 1000 1500 2000 2500 3000 3500 4000 4500 5000
EA+U_P EA and Unknowns with Definite and Probable cases
EA_P EA with Definite and Probable cases
EA+U EA and Unknowns with Definite cases
EA_D EA results with Definite Cases
OPR Previously found odds-ratios
Cases
Controls
chr18
53816275
chr16
50739582
17409572
chr11
17408630
114808902
chr10
6099045
chr9
32587450
chr6
32361265
150240076
chr5
35874575
111720761
chr4
111710169
chr3
114377568
chr1
67675516
BioVu

Networks



- You can expose the connections between data
  - Can visualize these connections
  - Analyze networks to expose bigger trends
  - http://diseasome.eu/map.html

Networks



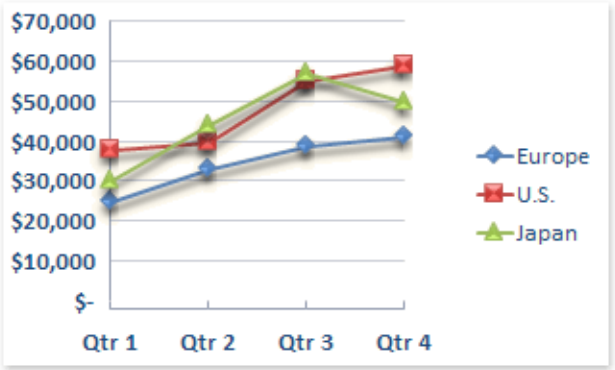
The Human Disease Network
Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007)

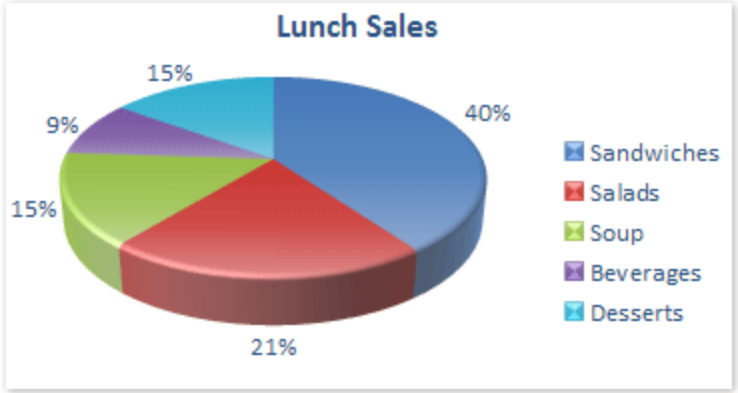
Data Visualization Software

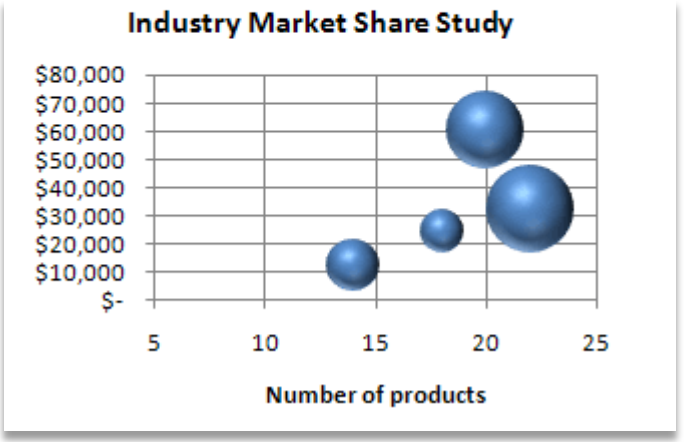
- Excel
- Pros
  - Nearly everyone has Microsoft Office
  - Bunch of automated tools for producing various kinds of plots
- Cons
  - Limited to the charts they offer
  - Repeating steps for same plots
  - CHART JUNK

Data Visualization Software

- What is chart junk?
  - Addition of unnecessary colors and/or shadow
  - Three dimensionality when the data only has two dimensions

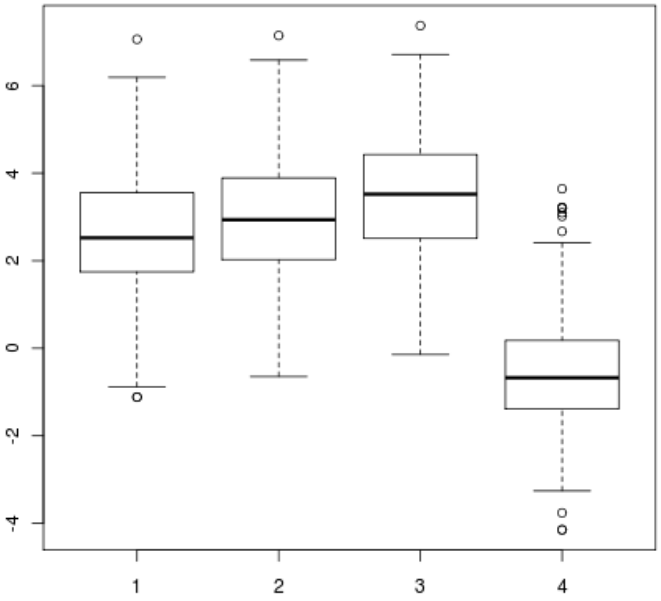


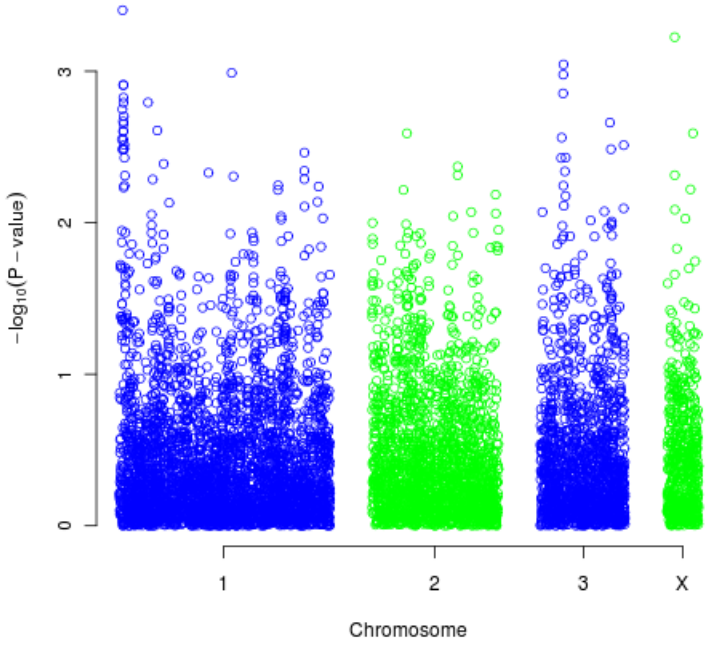




Data Visualization Software

- R
  - Language and environment for statistical computing and graphics
  - Individuals contribute packages to R
  - http://www.r-project.org/

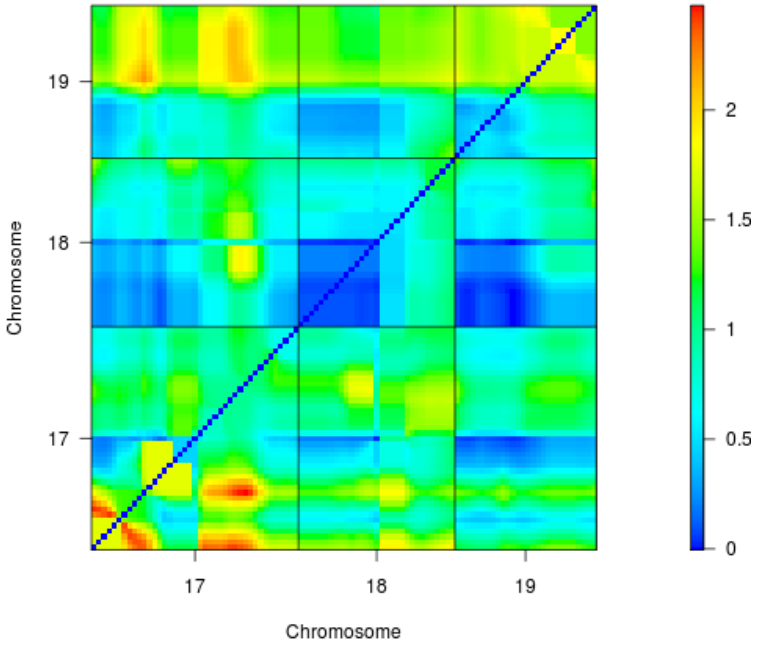


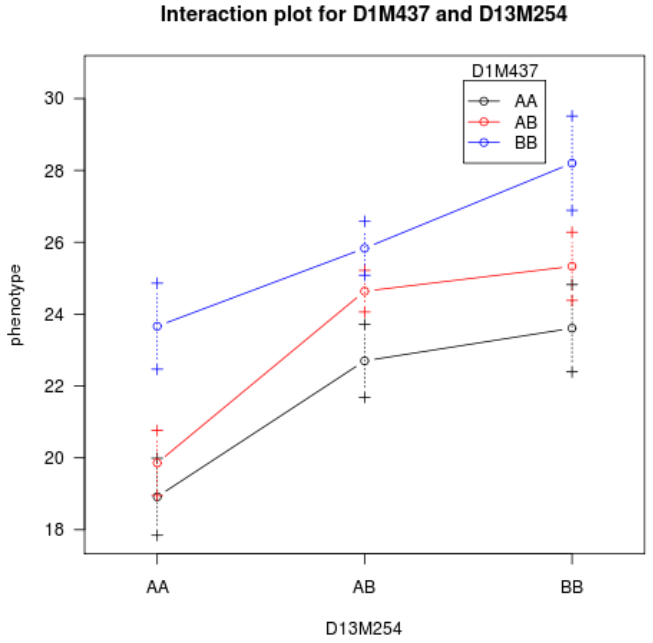


Graphics: http://rgm2.lab.nig.ac.jp/RGM2/

Data Visualization Software

- R
  - Language and environment for statistical computing and graphics
  - Individuals contribute packages to R
  - http://www.r-project.org/





Graphics: http://rgm2.lab.nig.ac.jp/RGM2/

Data Visualization Software

- R
  - Pros
    - Statistical and graphical options of every kind
    - Publication quality graphs
    - Once you have a script, easy to reuse and update
    - Statistics and graphics together
  - Cons
    - Information usually exists on how to make the plot, but finding that information can take a while!
    - Can be an investment in start up time to get a specific plot

Data Visualization Software

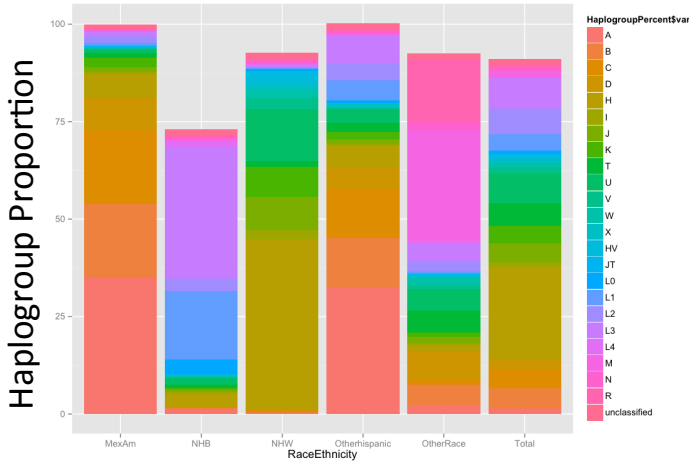
- R
  - Resources
    - The R project website
    - The R Book
    - "r project" web search
    - R Graph Galleries
      - Contains examples of plots as well as the code
    - R Graphical Manual: http://rgm3.lab.nig.ac.jp/RGM/
      - Also contains examples and code
    - R-Commander GUI interface for R

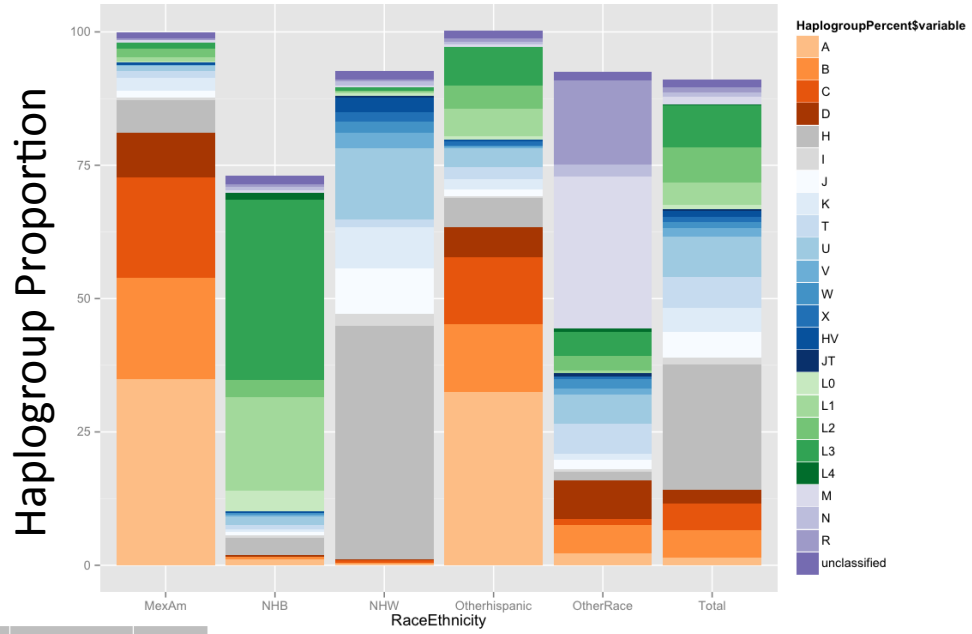
Data Visualization Software

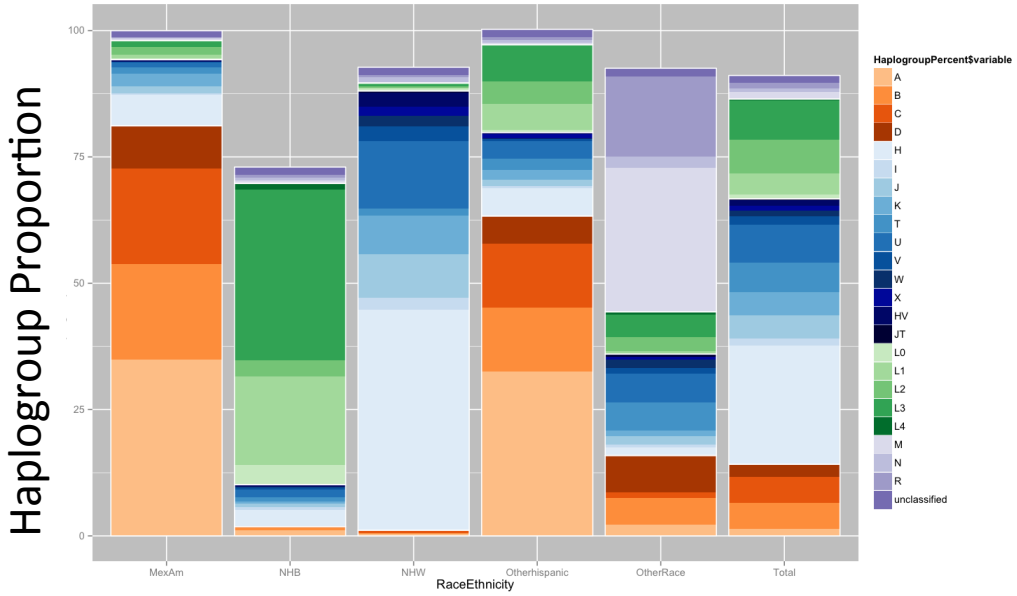
- GGPlot2
  - Plotting system for R
  - More automation for plotting data
    - Piece by piece
  - Clean graphical results
  - http://ggplot2.org/
  - Book: ggplot2: Elegant Graphics for Data Analysis

Data Visualization Software

- GGPlot2

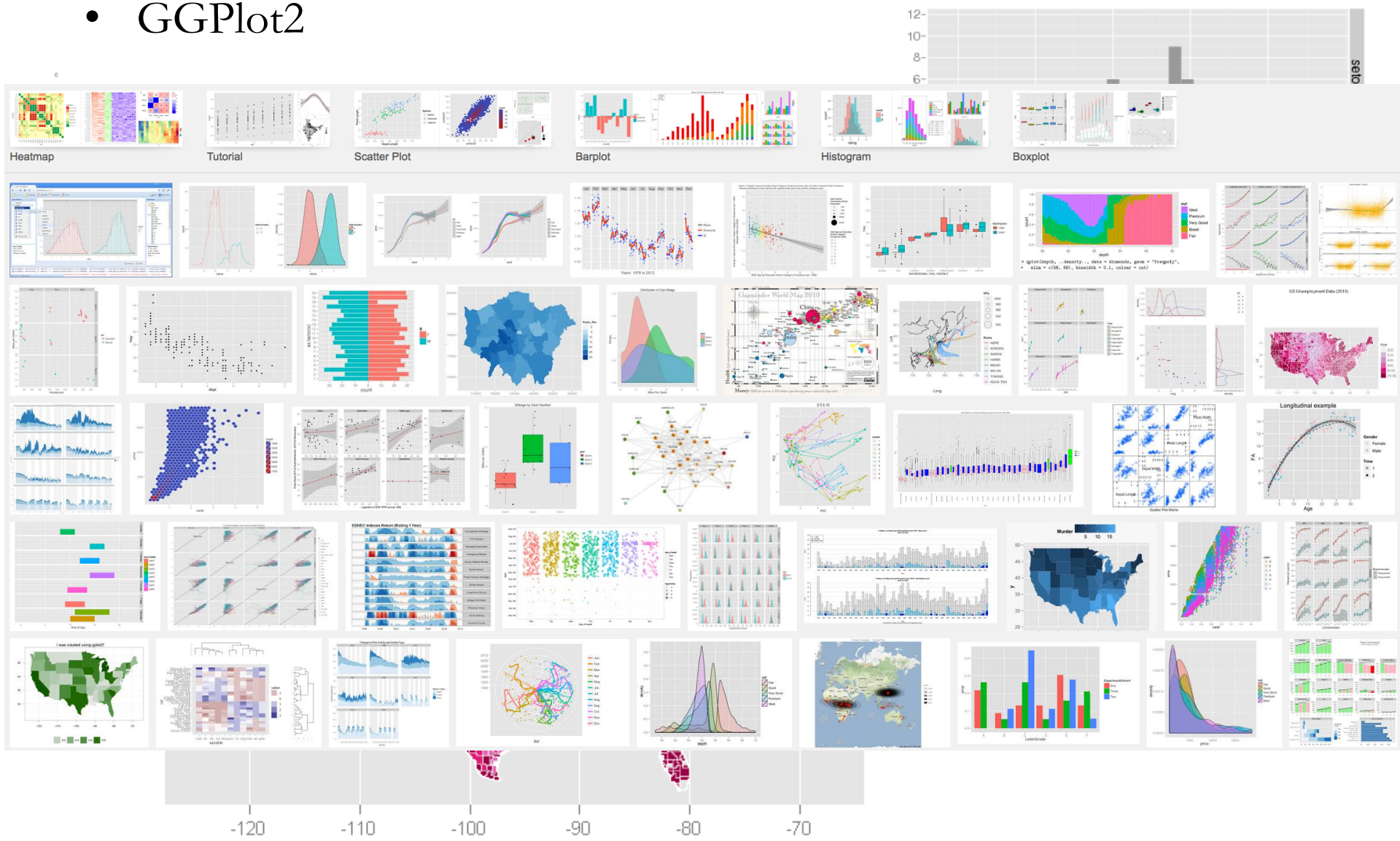






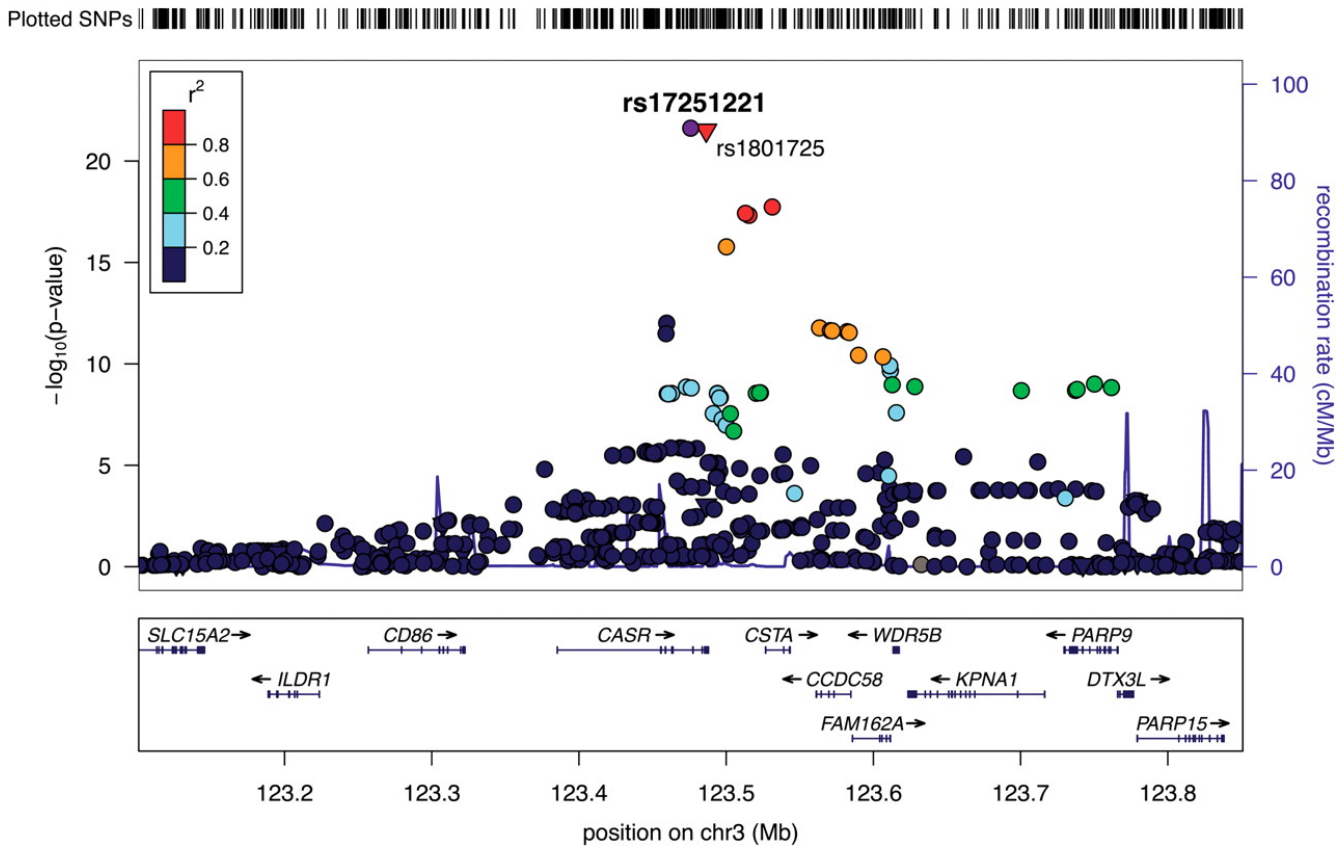
Data Visualization Software

- GGPlot2



Data Visualization Software

- Locus Zoom
  - Tool for plotting regional association results
  - http://csg.sph.umich.edu/locuszoom/faq.php

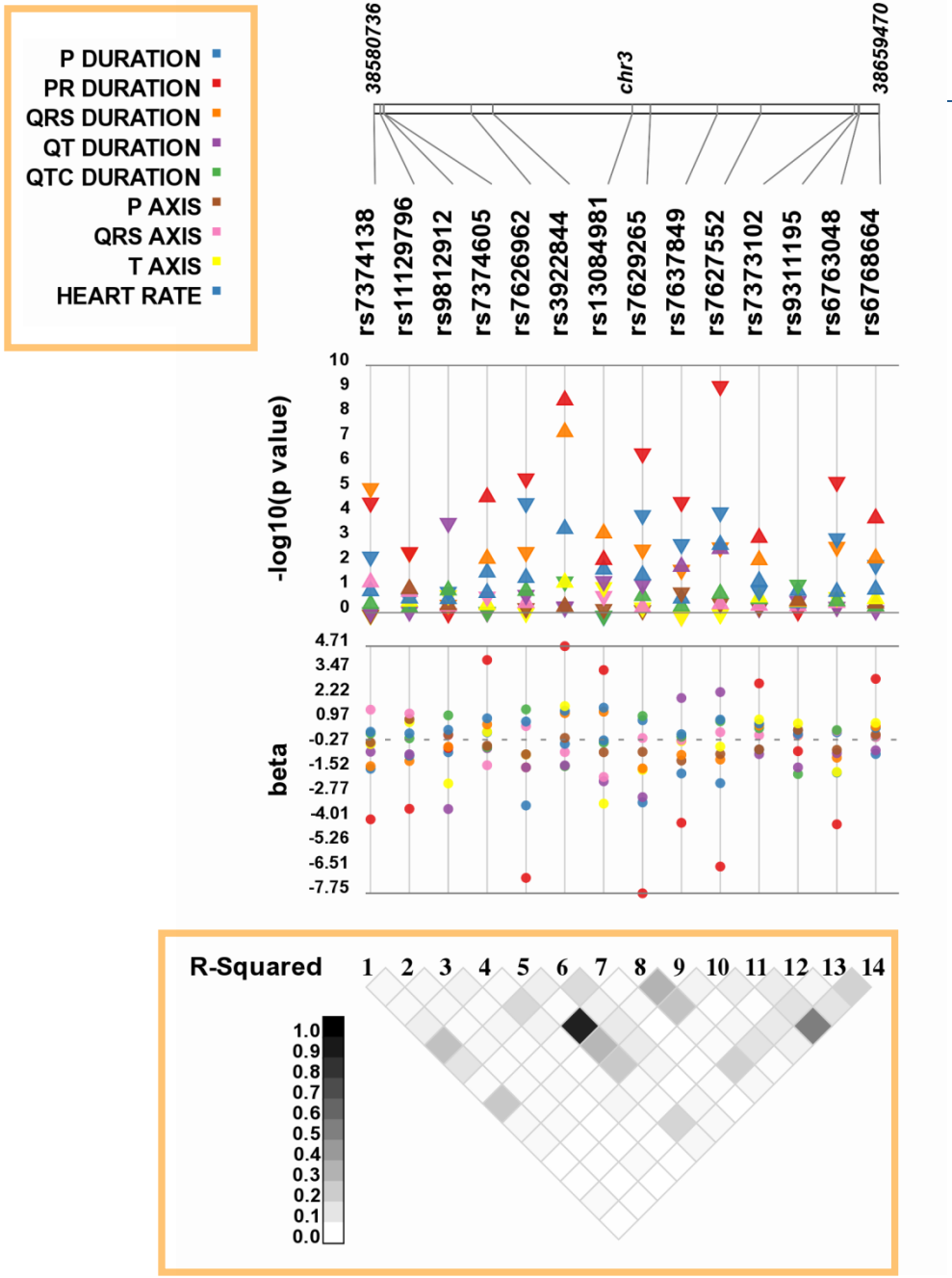


Graphic: O'Seaghdha et al. 2010

Data Visualization Software

- Synthesis-View software
  - Developed to visually synthesize multiple pieces of data
    - Stacked data-tracks
    - Multiple measurements within a single image
- Rapid visual comparison for results of various kinds
- Input requires formatted table
- For data characterizing < 100 SNPs
- http://visualization.ritchielab.psu.edu/

Synthesis View



P DURATION
PR DURATION
QRS DURATION
QT DURATION
QTC DURATION
P AXIS
QRS AXIS
T AXIS
HEART RATE

SCN5A variation is associated with electrocardiographic traits in the Jackson Heart Study, Jeff et al., Circulation and Cardiovascular Genetics, 2010

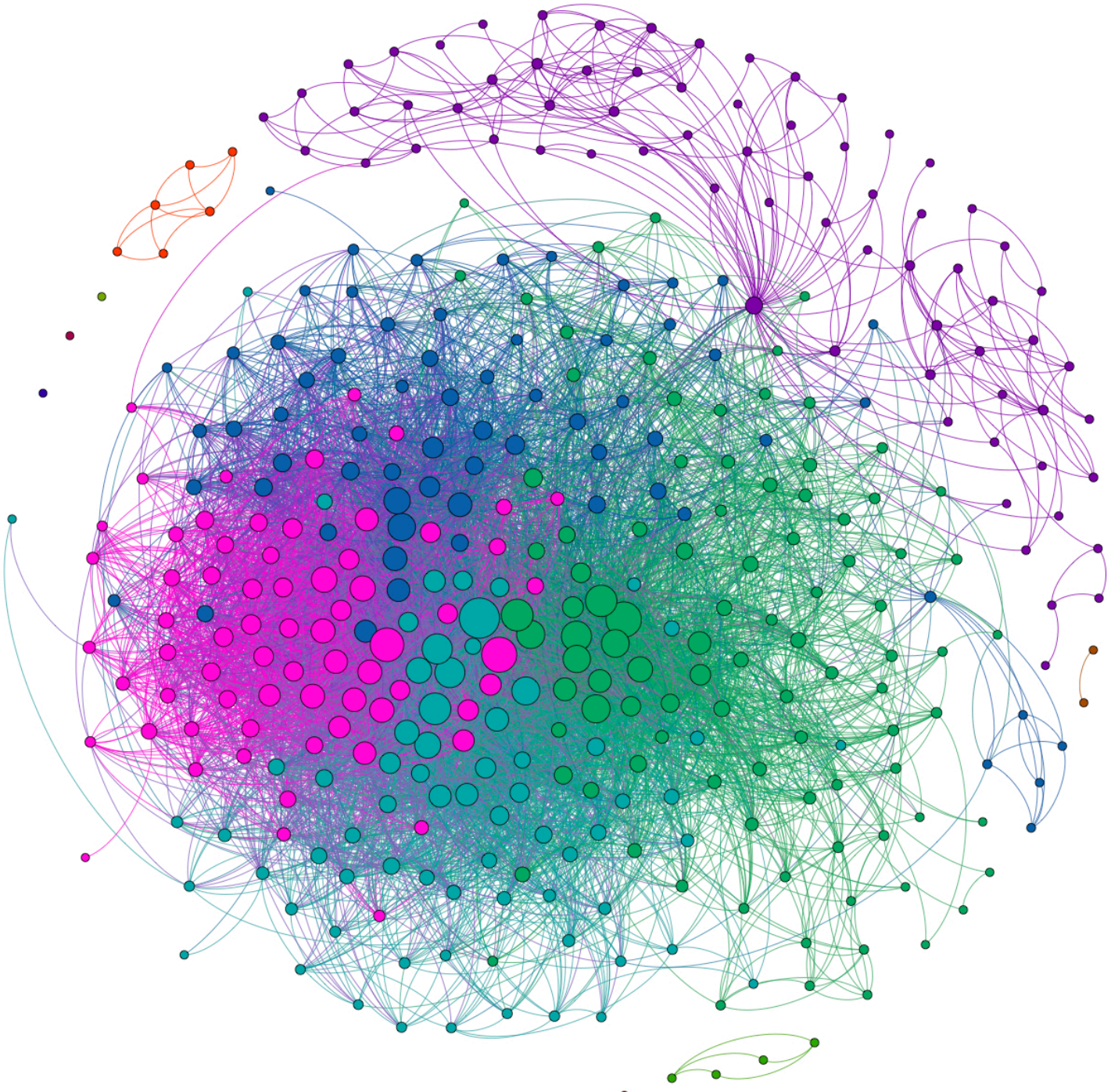
Data Visualization Software

- Interactive Plots
  - Data Driven Documents: D3
  - Java Script Library
  - http://d3js.org/

Data-Driven Documents

ıre

- Net
  - (
  - h
  - h



be

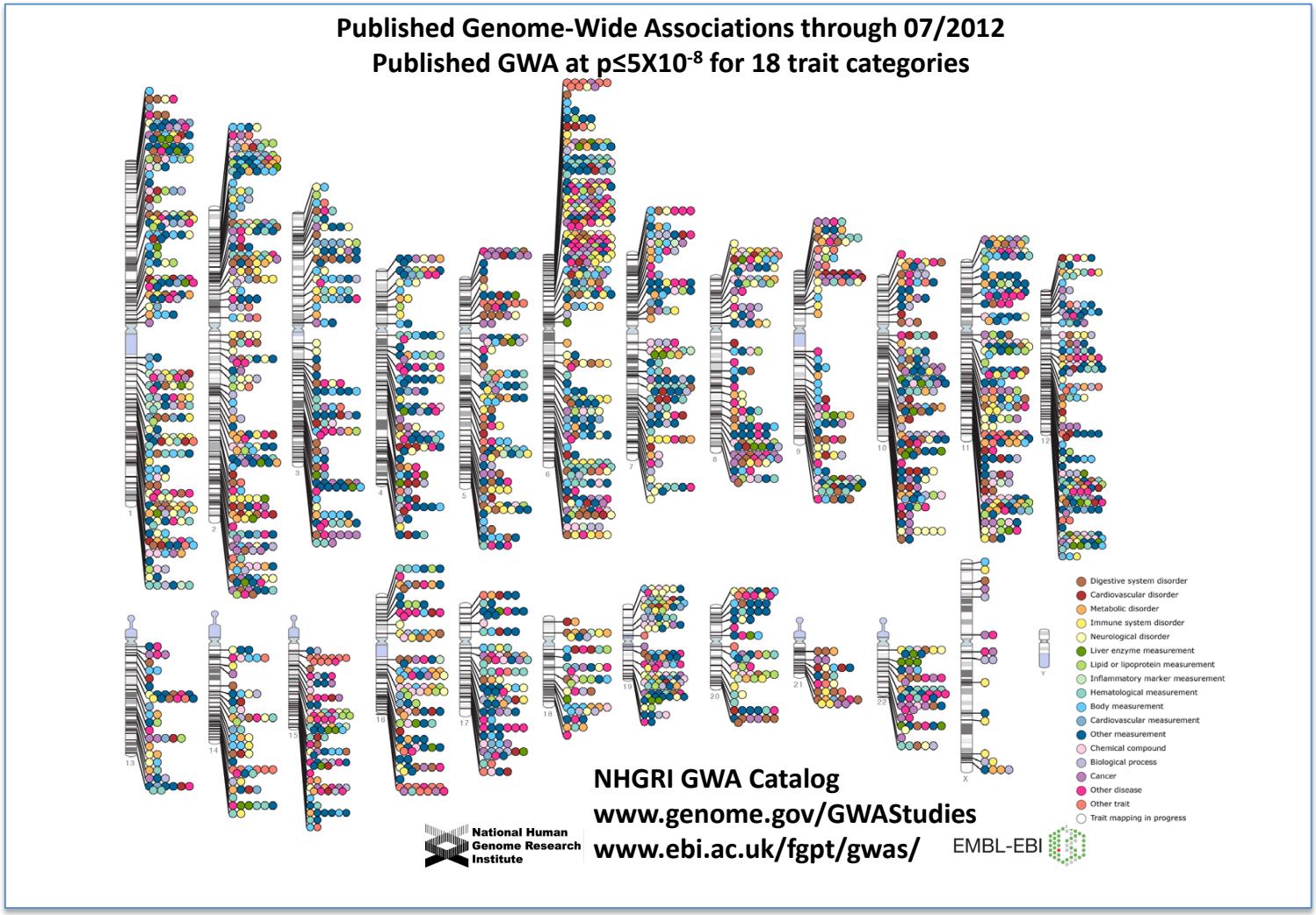
nsortium

The Open

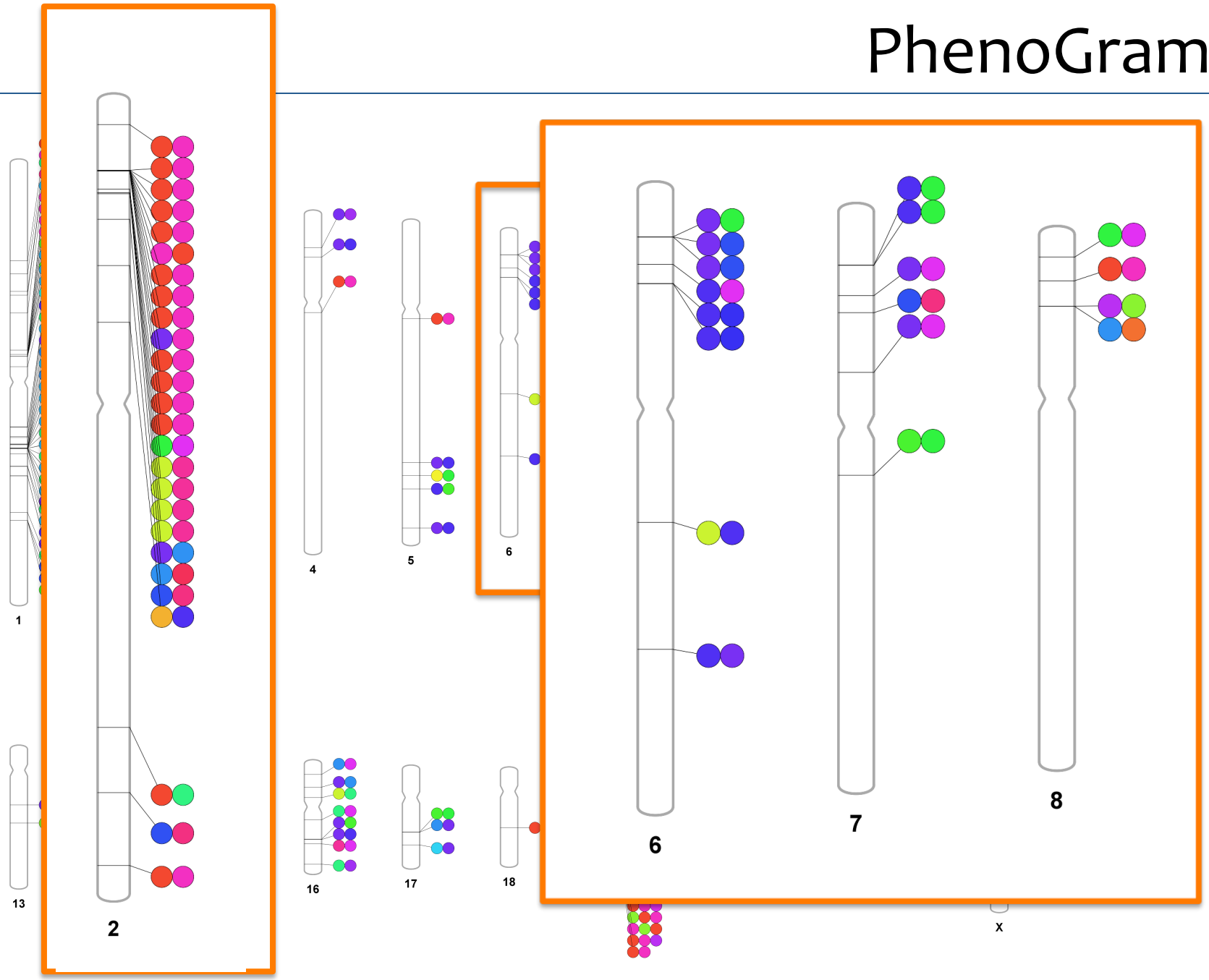
Gephi is an inter
exploration platf
and complex sys
graphs.

Runs on Windows
open-source and

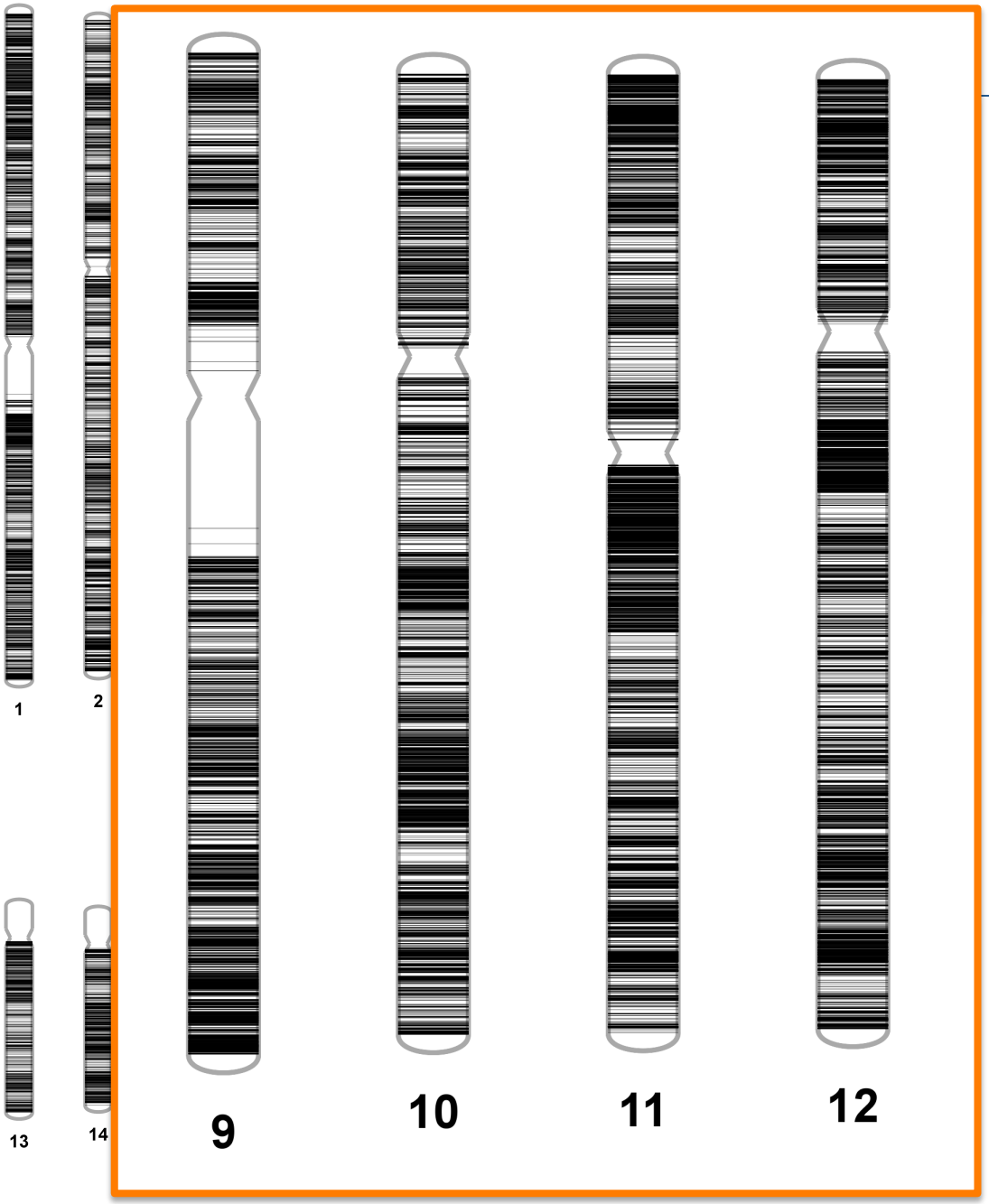
NHGRI GWAS Catalog Plot



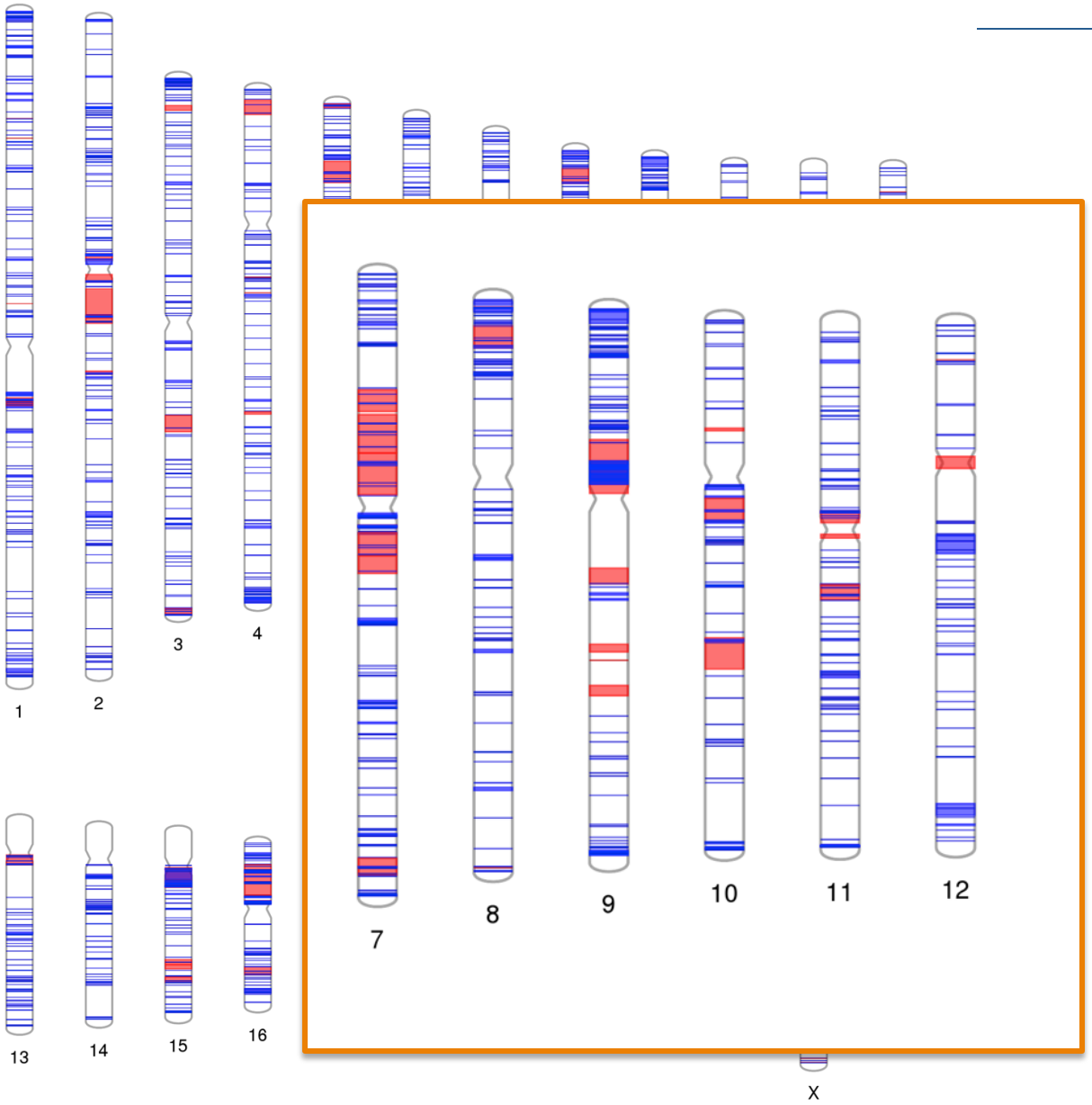
PhenoGram



PhenoGram



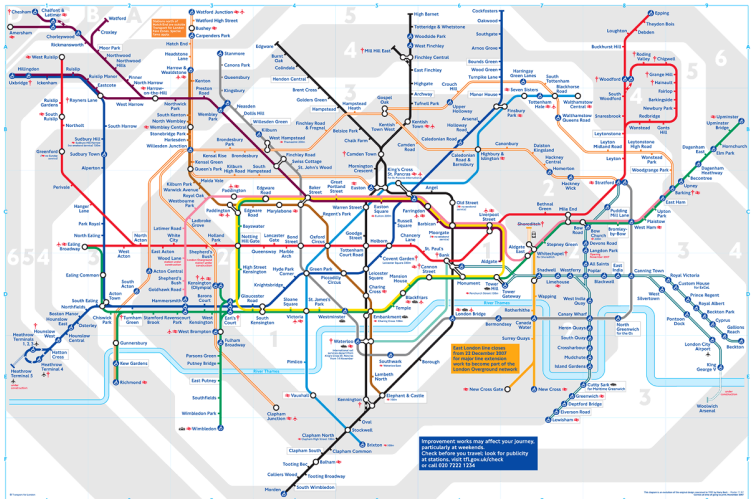
PhenoGram



Data Visualization Software

- Additional software packages
  - STATA
  - Matlab
  - SAS/Graph software

Data Visualization



Data Visualization

- Suggested Reading
  - Edward Tufte
  - http://flowingdata.com/
  - http://infosthetics.com/
  - http://www.visualcomplexity.com/vc/

Data Visualization

- Questions?