Alexander Lex
@alexander_lex
http://alexander-lex.net

THE UNIVERSITY OF UTAH

Visualizing OMICS Data

visualization design lab

SCI
www.sci.utah.edu

visualization
design lab

http://vdl.sci.utah.edu/

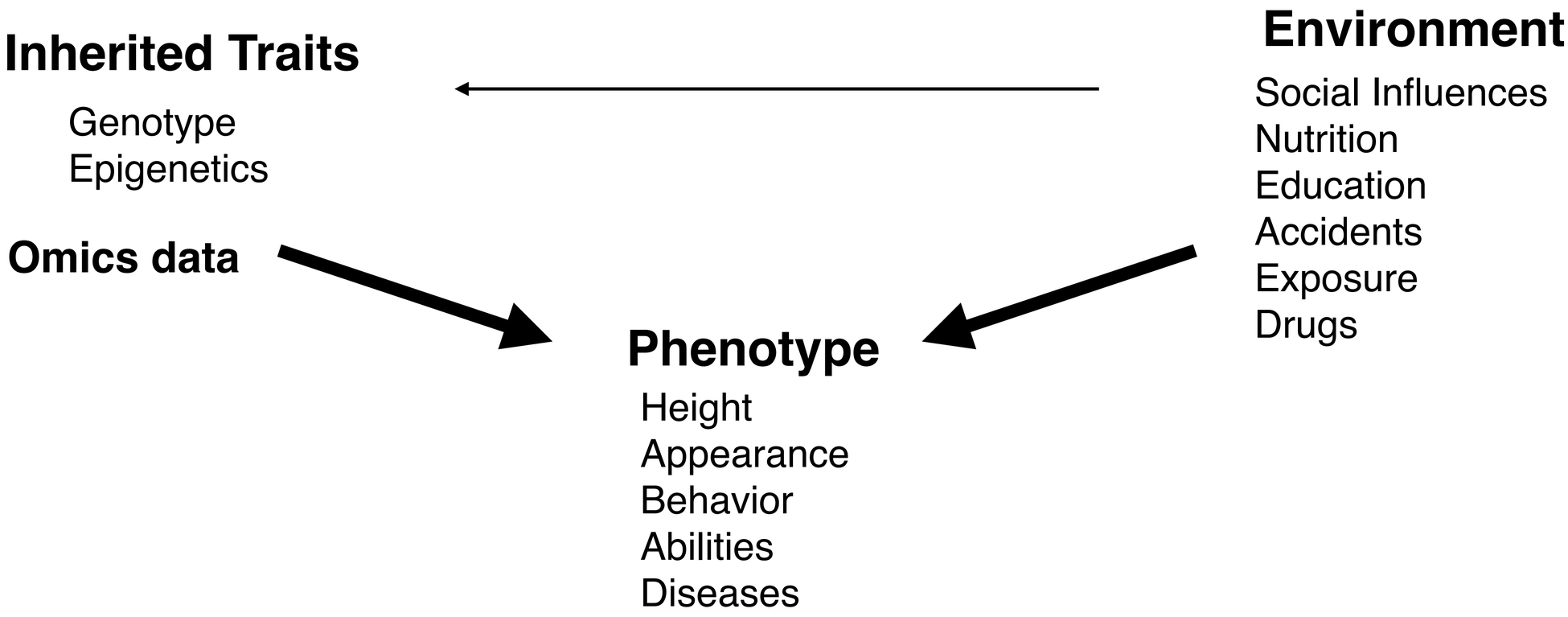
TOPIC: OMICS VISUALIZATION

on Genome Coordinates (structural perspective)

in Tabular Form (experiments, multiple conditions)

on Networks (functional perspective)

Genetics & Phenotype



Omics

Shorthand for different branches of biology:

Gen**omics** – study of gene sequences

Prote**omics** – study of proteins

Metabol**omics** – study of metabolism

Transcript**omics** – study of transcription processes

Epigen**omics** – study of the supporting structures of the DNA

etc..

Why is Understanding Omics Important?

Understanding Fundamentals in Biology

Disease Prevention

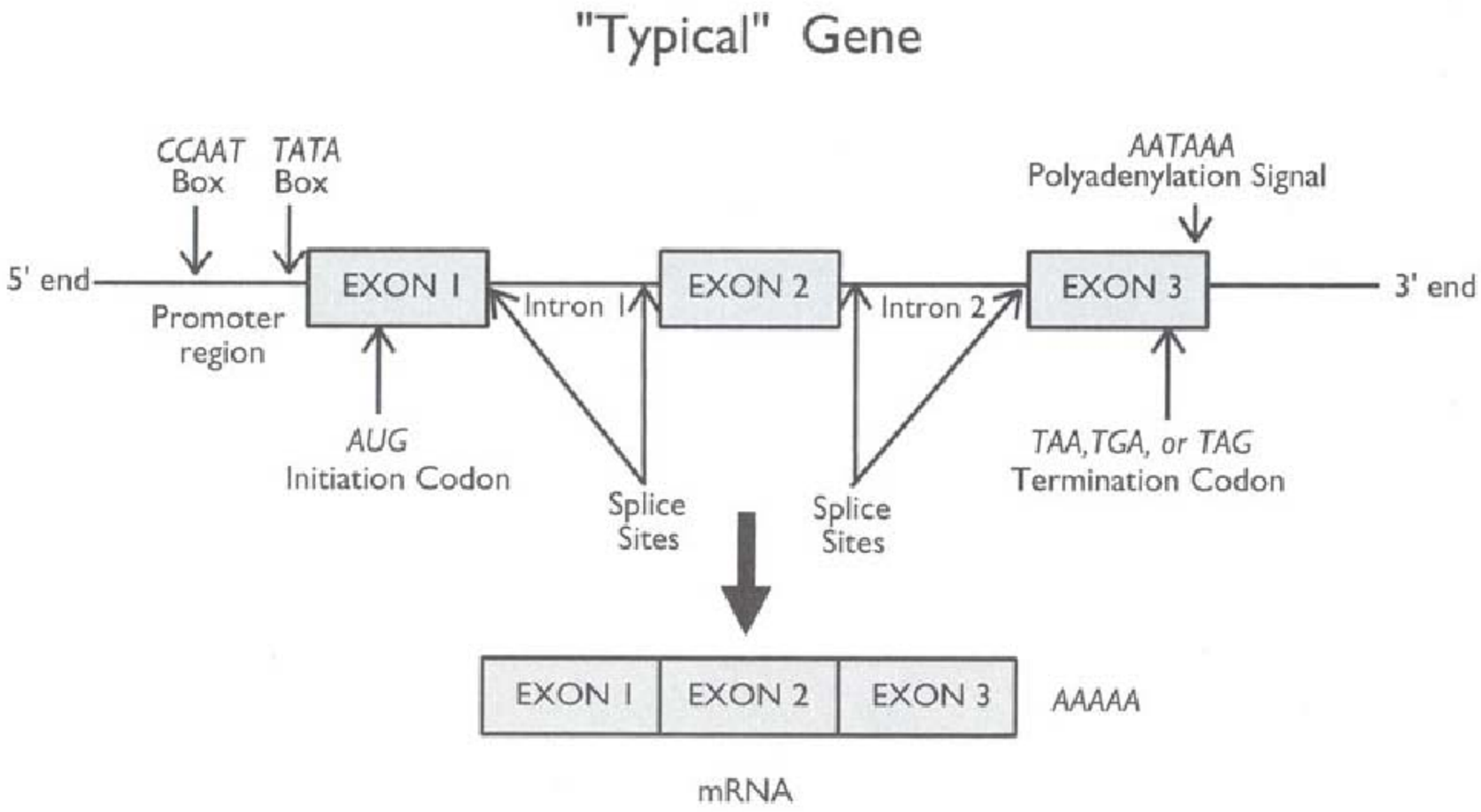
Targeted Diagnosis (BioMarkers)

Personalized Medicine

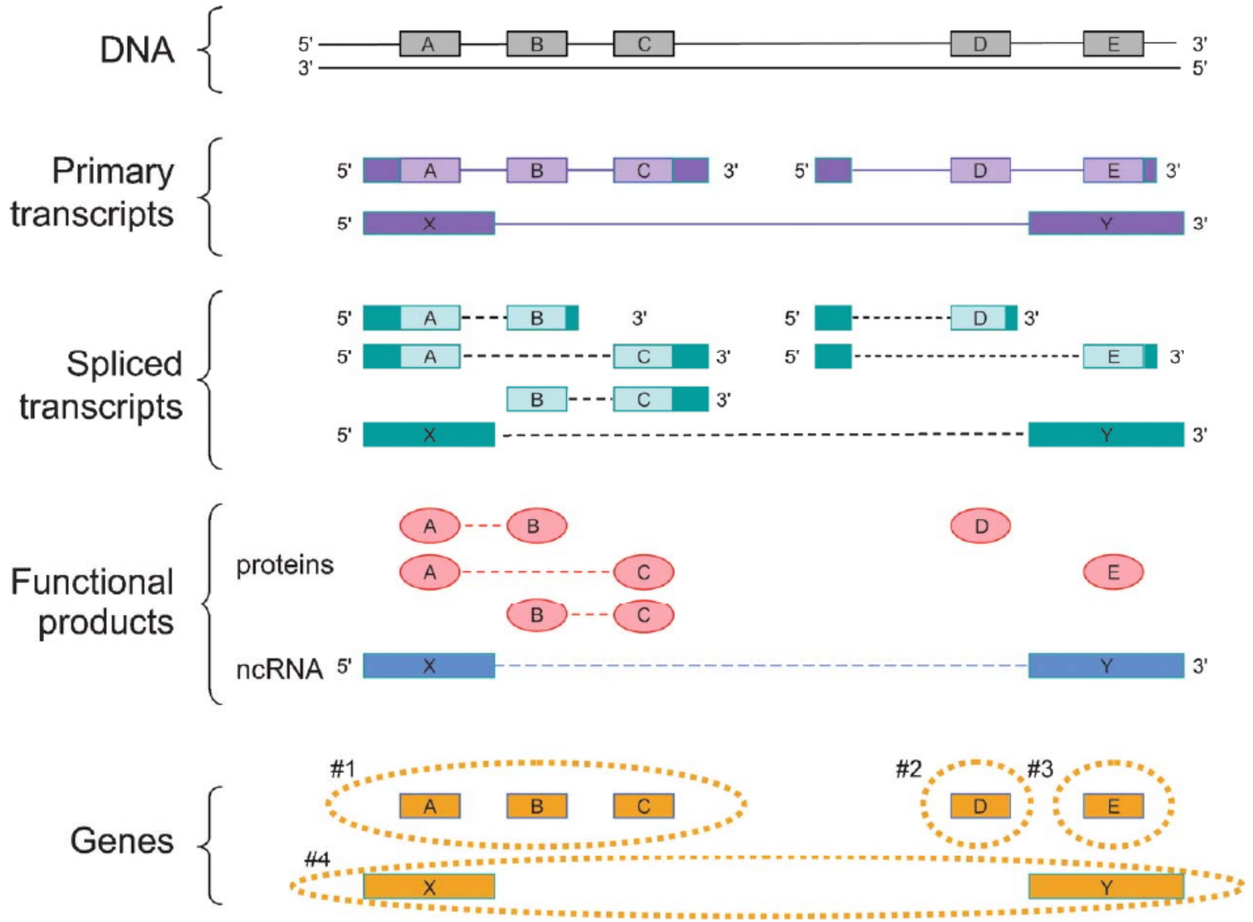
Drug Development

Targeted Modification of Organisms

Genes


"Typical" Gene
CCAAT Box
TATA Box
AATAAA Polyadenylation Signal
5' end
EXON 1
Intron 1
EXON 2
Intron 2
EXON 3
3' end
Promoter region
AUG Initiation Codon
Splice Sites
Splice Sites
TAA, TGA, or TAG Termination Codon
EXON 1
EXON 2
EXON 3
AAAAA
mRNA

Modern Gene


DNA
5'
3'
A
B
C
D
E
3'
5'
Primary transcripts
X
Y
Spliced transcripts
Functional products
proteins
ncRNA
Genes
#1
#2
#3
#4

From DNA to Functional Protein

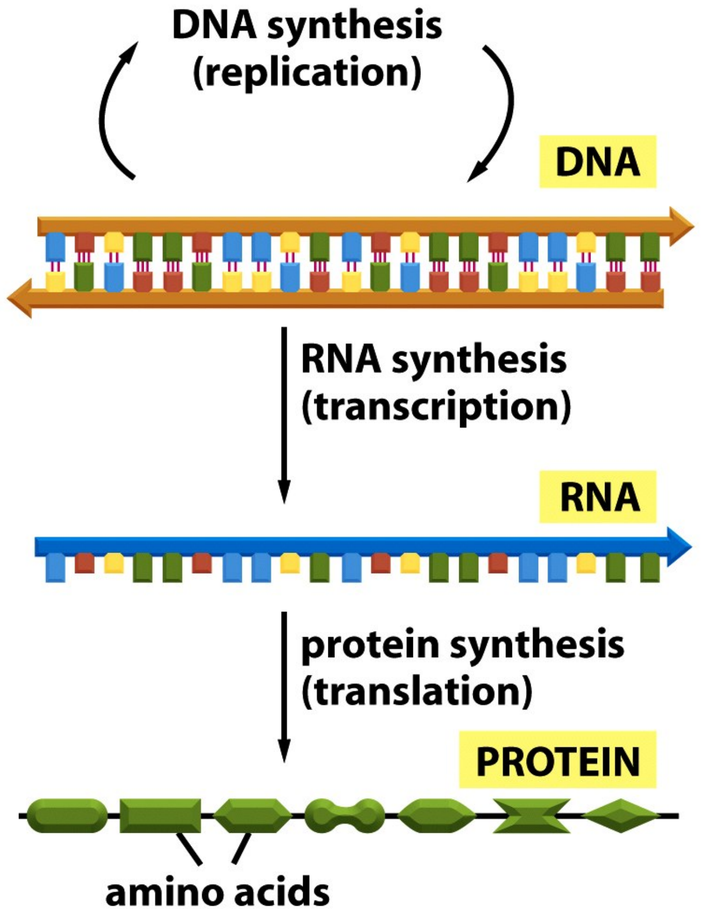
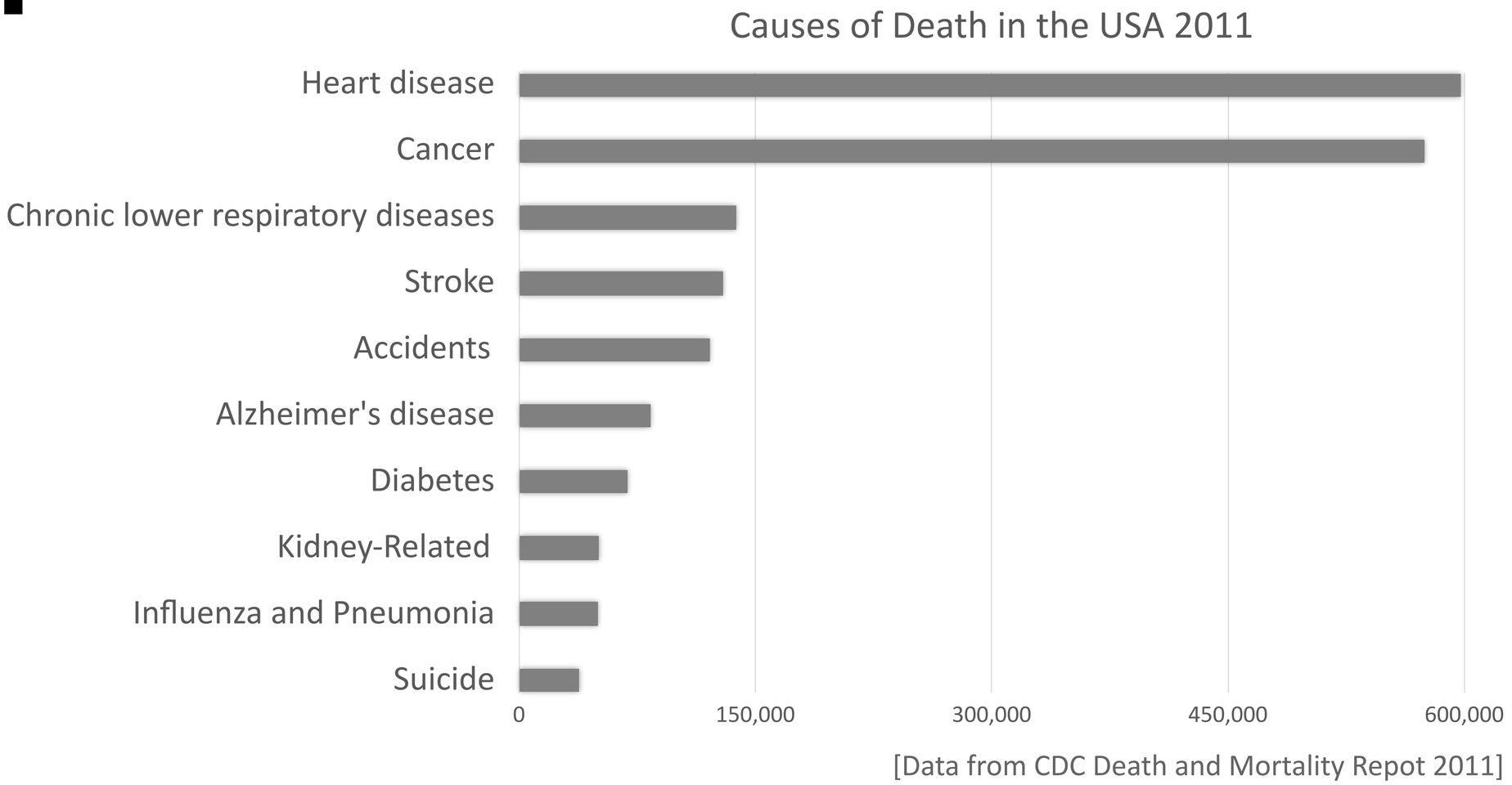


Figure 1-4 Molecular Biology of the Cell 5/e (© Garland Science 2008)

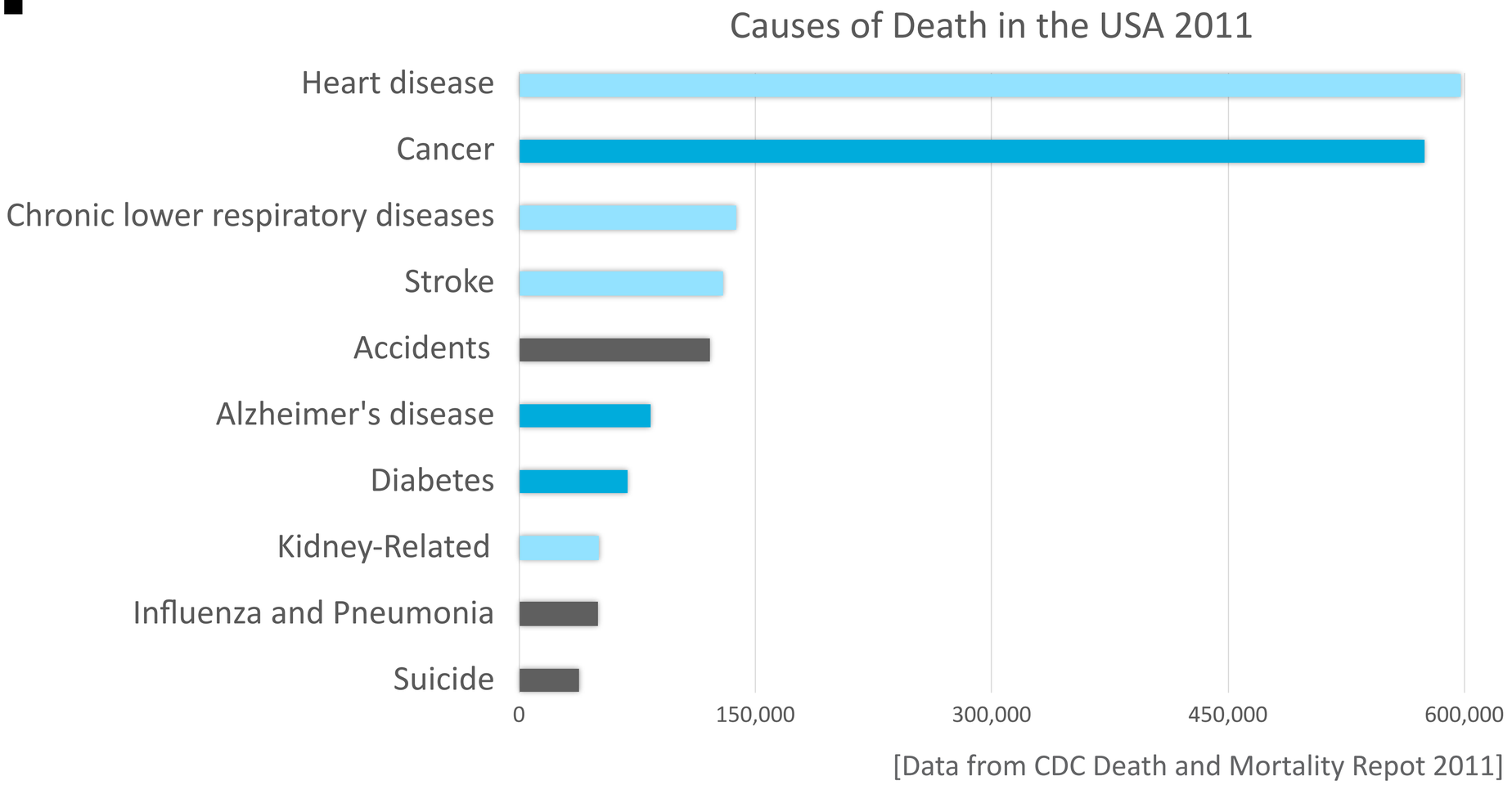
Protein Functions

structural support
storage of amino acids
transport of other substances
coordination of an organism’s activities
response of cell to chemical stimuli
movement
protection against disease
selective acceleration of chemical reactions

Why is Understanding Omics Important?



Why is Understanding Omics Important?

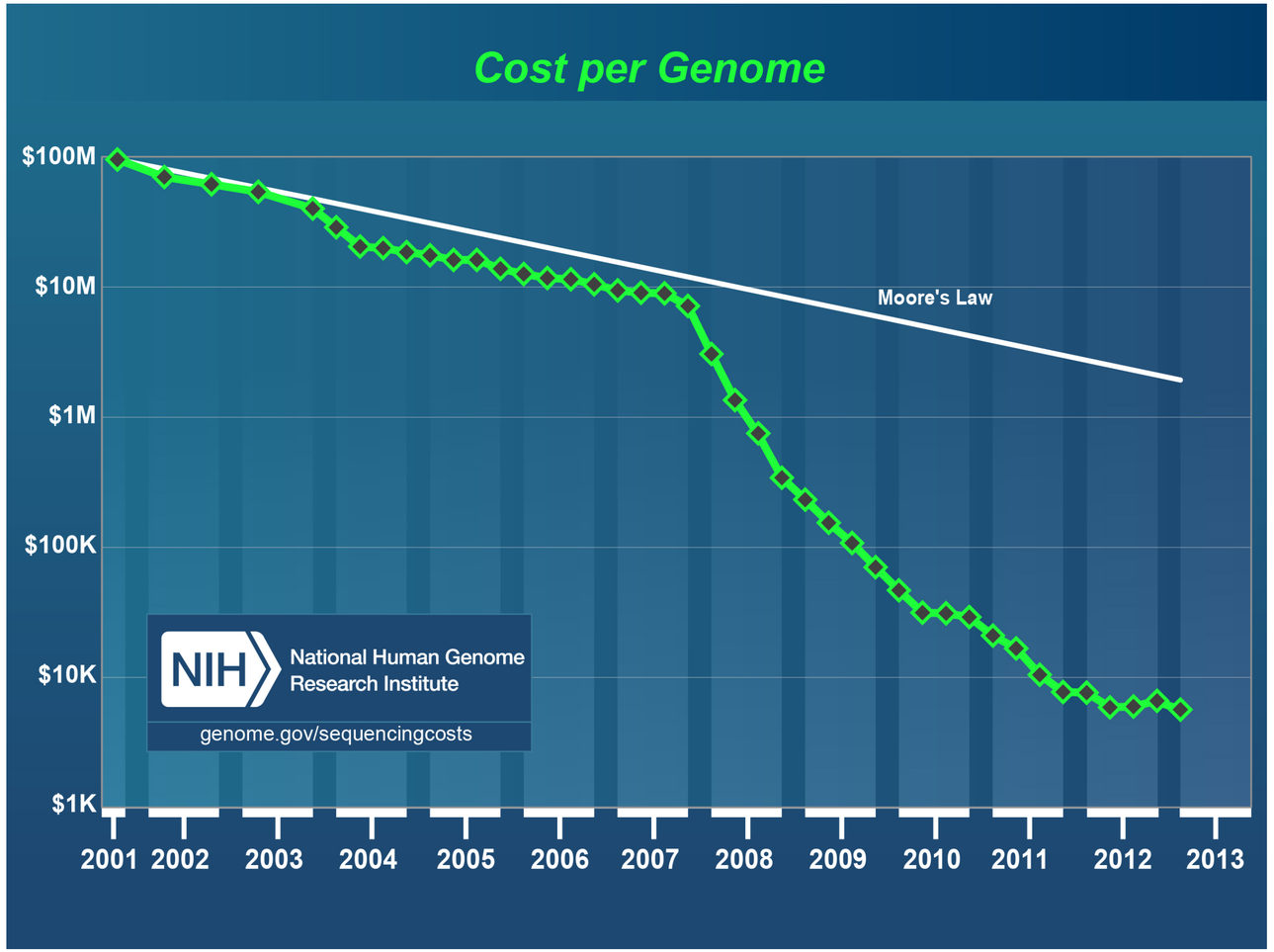


Why is Vis for Omics Data important?

Biology is in the middle of a revolution!

- Transformation from a wet-lab/experimental to computational science
- Challenge in MB is shifting from *Data Acquisition* to *Data Processing & Analysis*

Why is Vis for Omics Data important?



What does this mean?

We can now do very large experiments

The Cancer Genome Atlas

Understanding genomics to improve cancer care

Why is the Analysis Hard?

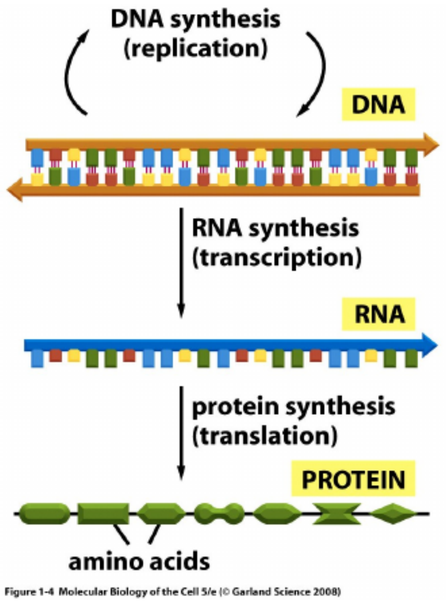
20,000 protein coding genes (1.5% of the genome)

3 billion basepairs

Gene -> Protein -> Function

Each of these steps is influenced by many processes (all the OMICSes)!

Very complex interplay of functional aspects



Main Types of Omics Data

Sequence Aligned -> Genome Browsers

Experiments vs Activity (Abundance) -> Tabular Data

Systems Biology -> Omics Data Overlaid on Networks

Sequence Aligned Data

Example Tasks:

Identify Structural Variations

Browsing annotations and mapped experimental data

Comparing sequences

Data Acquisition via **Genome Sequencing**

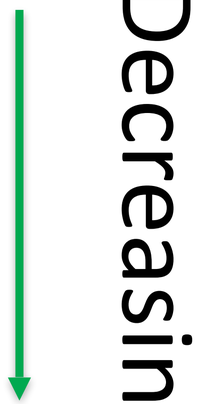
Types of Structural Variability

Chromosomal alterations

Copy-number variation

Mutations

SNPs

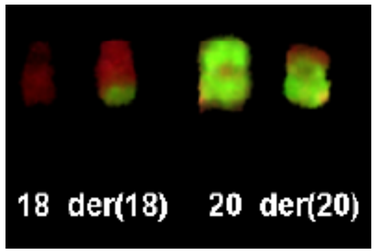


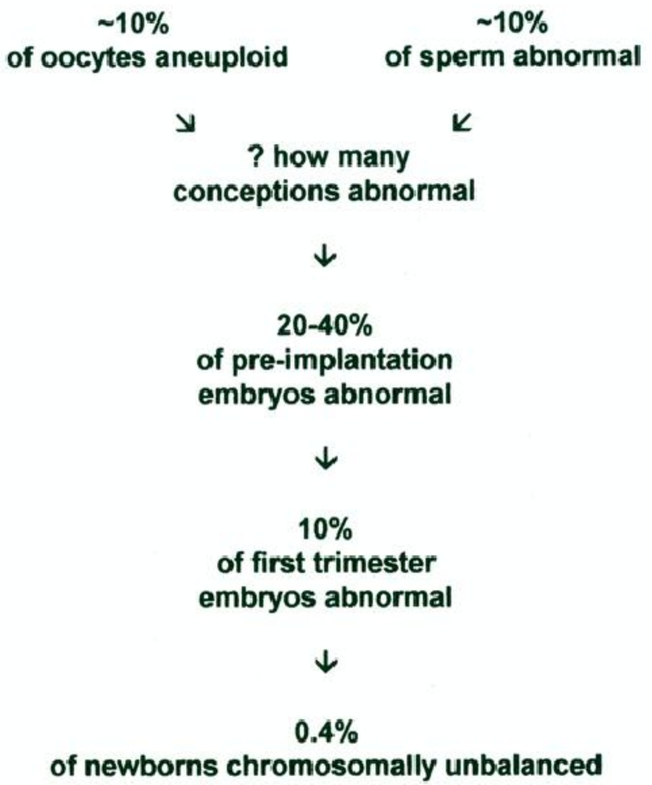
Chromosomal Alterations

Missing or duplicate chromosomes

Trisomy 21 (Down Syndrome) most common (95%)

Other chromosomal defects are very common but mostly fatal for embryos





Copy Number Variaton (CNV)

Length of > 1 kilobase

No clear boundary to chromosomal alterations

Sequence can be

- amplified (multiple copies), or
- deleted on one (heterozygous) or both (homozygous) chromosomes

Leads to increased/decreased activity or loss of function

Mutations

Point Mutations

- Different base in a single location
- Silent mutations: which code for the same (or a sufficiently similar) amino acid
- Missense mutations: which code for a different amino acid
- Nonsense mutations: which code for a stop and can truncate the protein

Mutations

Insertions

- One or more extra bases

Deletions

- Missing one or more bases

“InDels”

Inversion

Common, about 30,000 in healthy human

Mutations

Can have no effect on the phenotype

Can lead to severe diseases

Cystic fibrosis -> CFTR mutation

IDH1 -> Brain Cancer, BRCA1/2 -> Breast Cancer

Single Nucleotide Polymorphisms

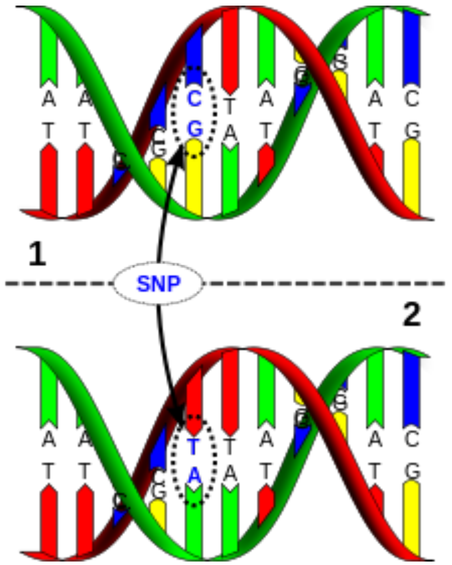
Occur on average every 1000-2000 basepairs

Responsible for the natural phenotypical variation between organisms

Only considered SNP when variations are common in the population (else point mutation)

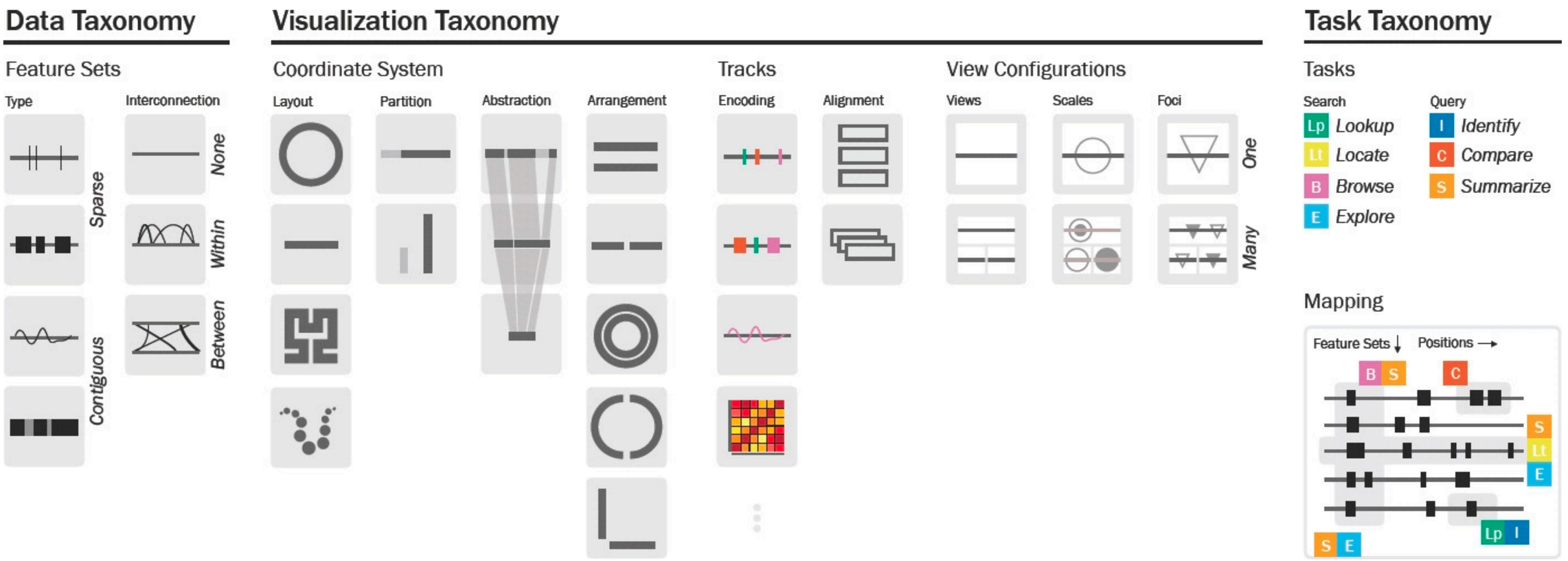
Can play role in diseases

e.g., Alzheimer



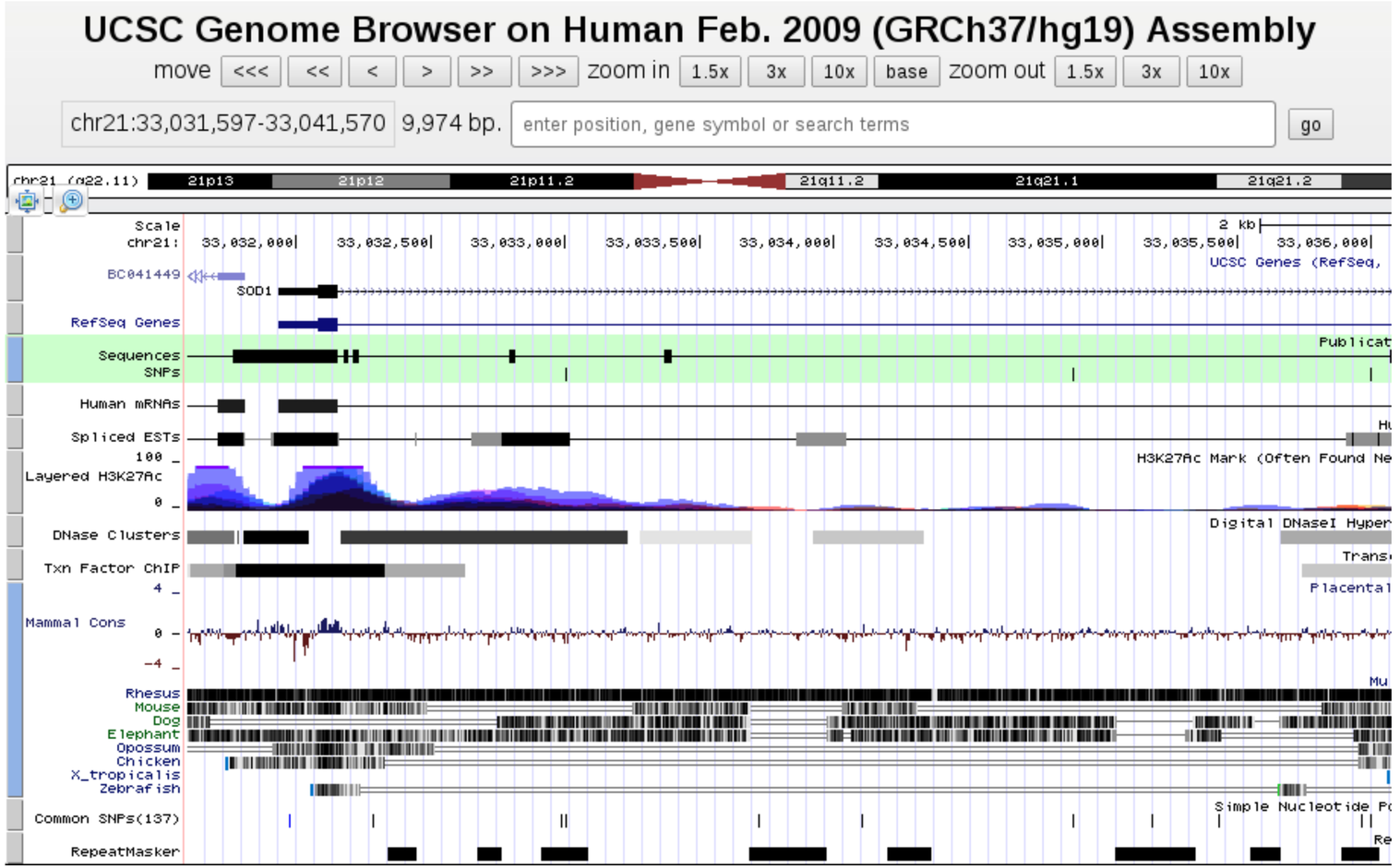
Browsing Genomes

- Alignment of various data by sequence
  - Annotations
    - Genes or other active regions
  - Gene Expression Data
  - Genotype Variation
- Based on “Tracks”

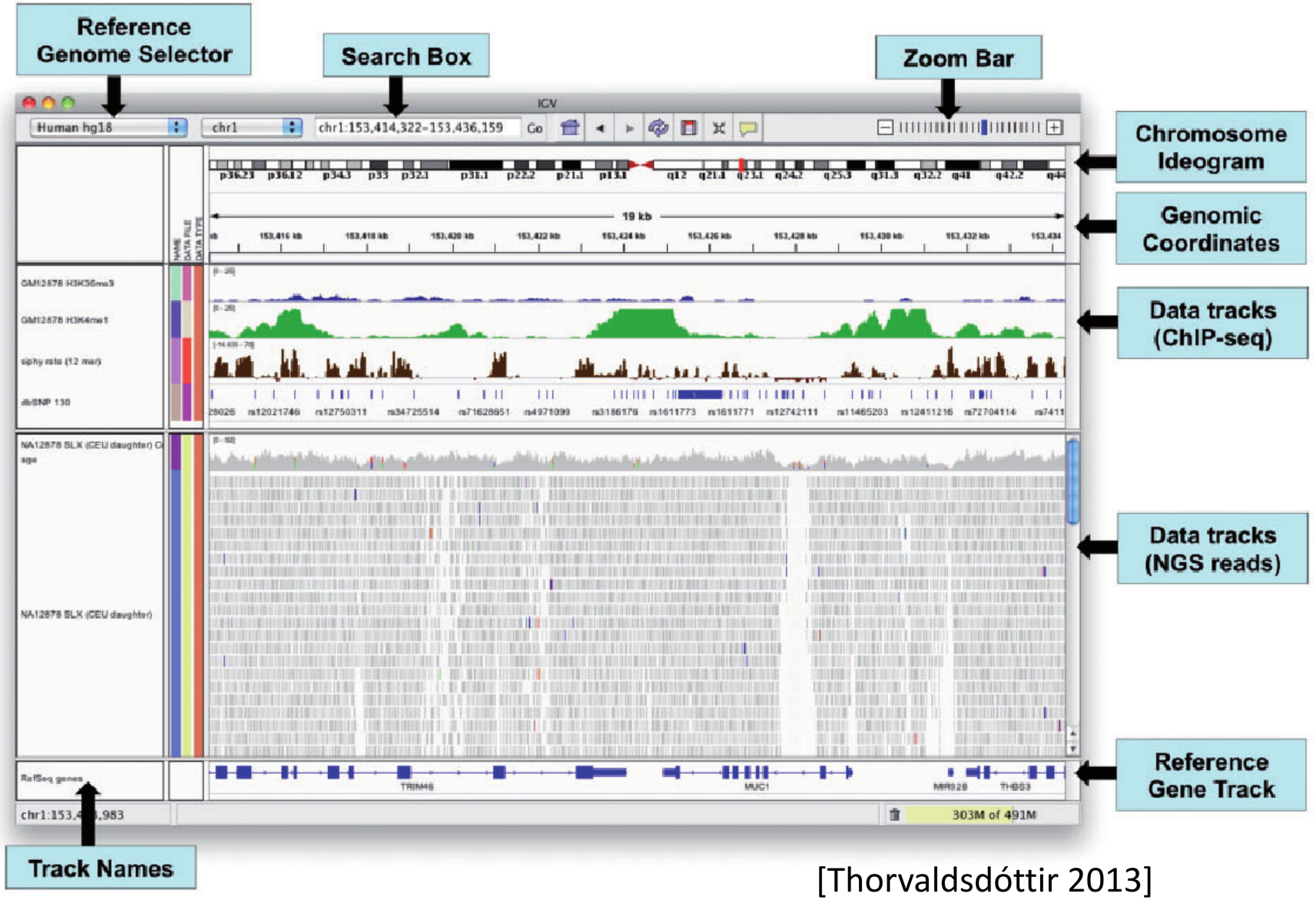


S Nusrat, T Harbig, N Gehlenborg. “Tasks, Techniques, and Tools for Genomic Data Visualization” *Computer Graphics Forum*

UCSC Genome Browser



Integrative Genome Viewer (IGV)



[Thorvaldsdóttir 2013]

Genome Browsers

Good if you need high-res data

- access to single nucleotides

Good if chromosome-orientation makes sense

Problematic: functional relationships don’t correspond to sequence

Sequence is primarily interesting for evolution and structure, not for function

Different Ways to Depict Translocations

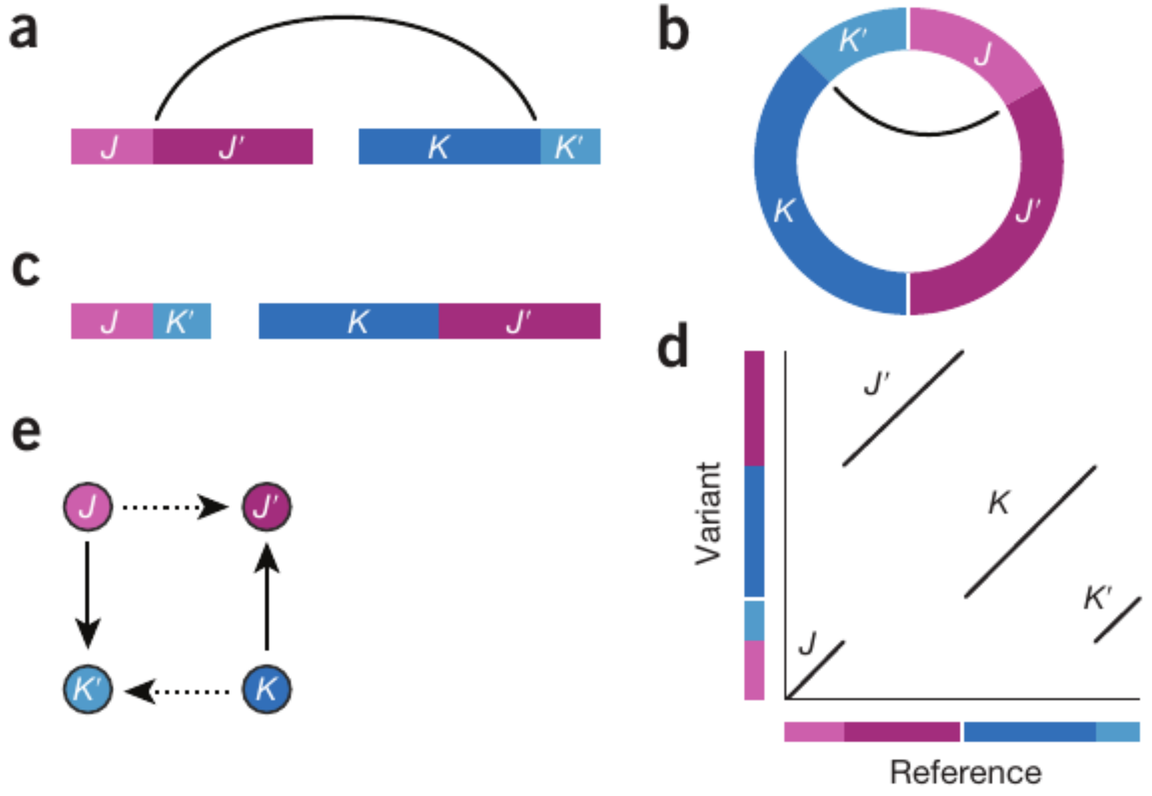
a. Translocation shown by arc

b. Translocation shown by arc, circular layout

c. Translocation actually shown, color coding the chromosome

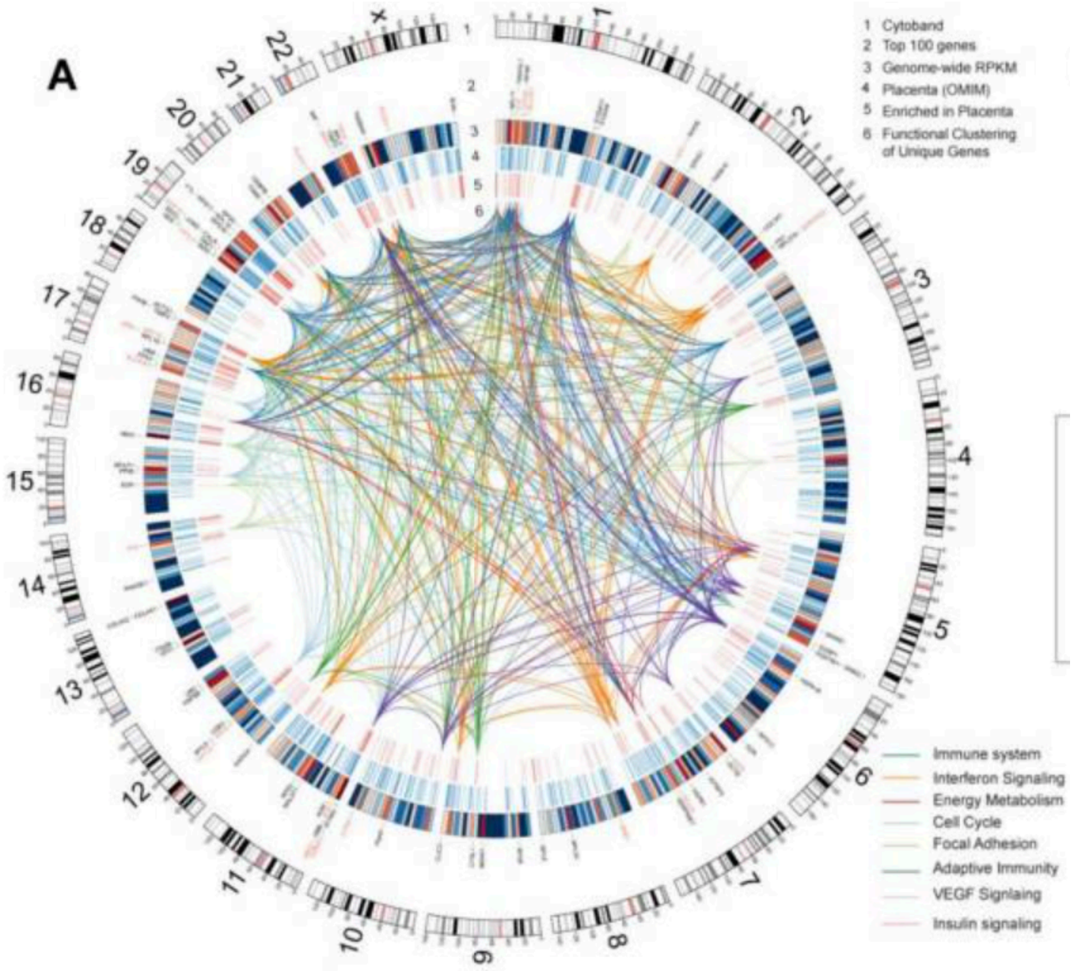
d. Translocation in dot plot

e. Non-linear encoding using graph – dashed original, solid changed



[Nielsen 2012]

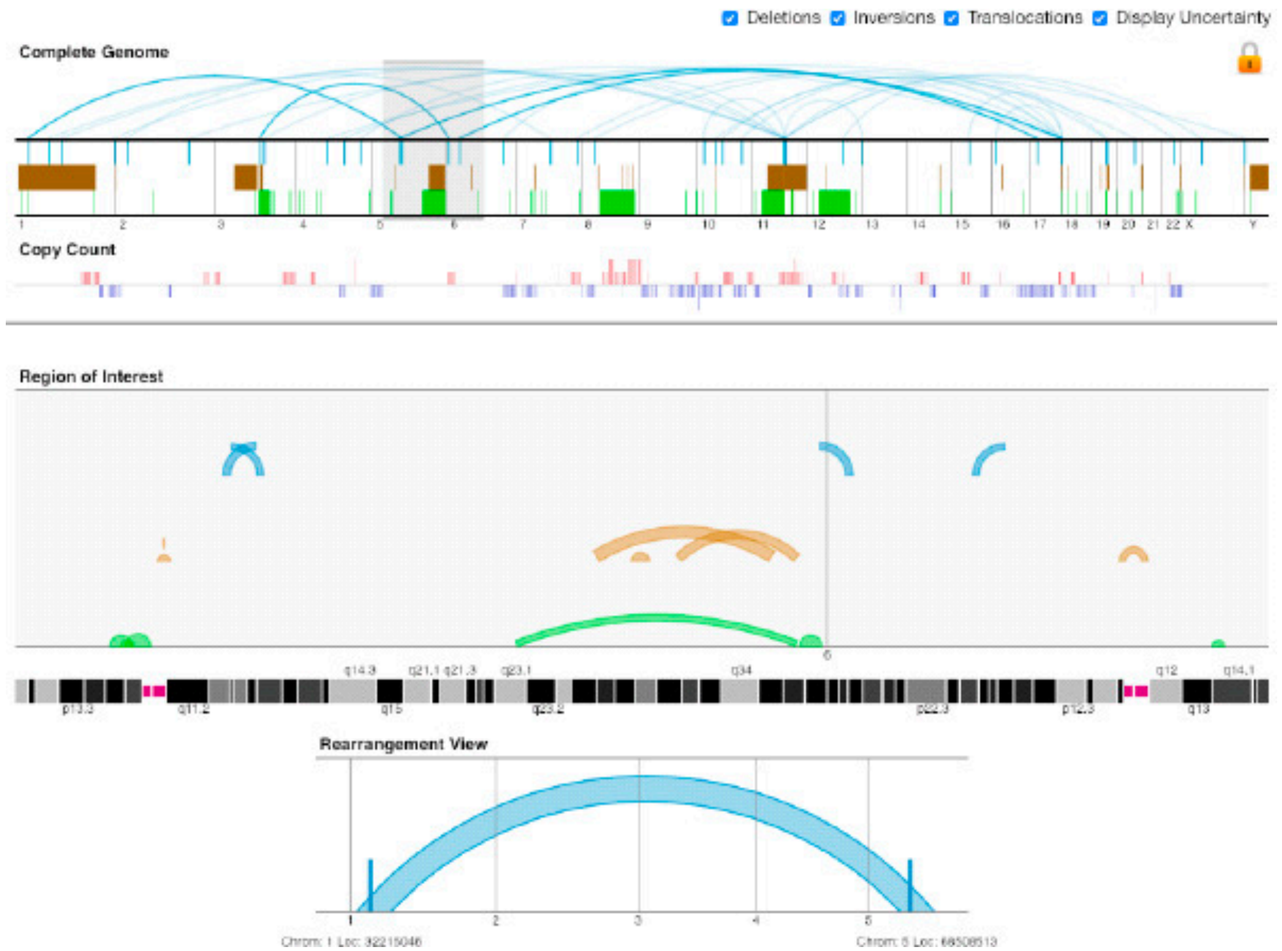
Circular Layout – Circos



http://circos.ca/

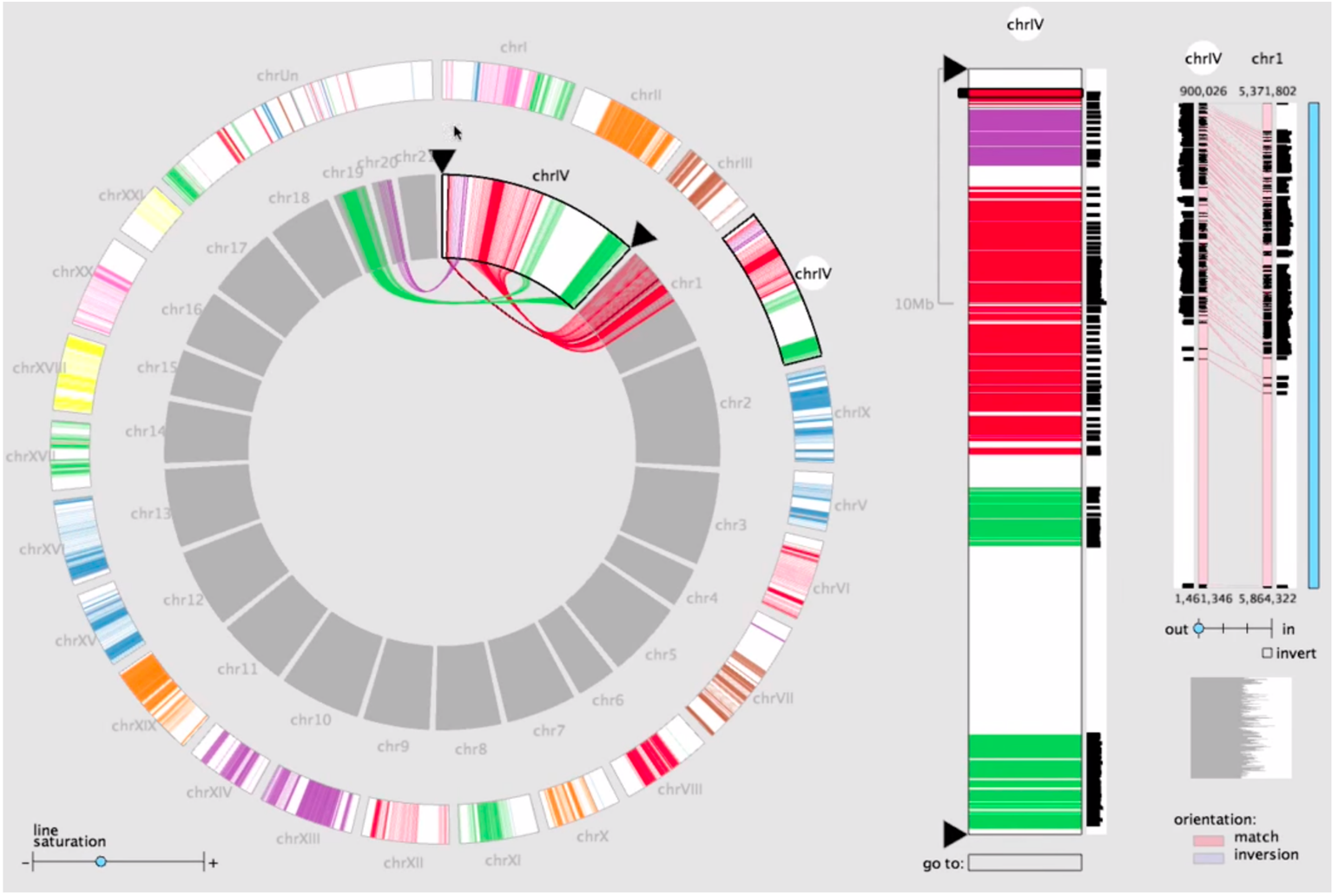
[Krzywinski 2009]

Linear Layout – Gremlin



O'Brien T. M., Ritz A. M., Raphael B. J., Laidlaw D. H.: Gremlin: an interactive visualization model for analyzing genomic rearrangements. *IEEE Trans. Vis. Comput. Graph.*

MizBee – Comparison Between Species



[Meyer 2009]

Tabular Data

Abundance of Some Omics Data Type

Gene Expression

Protein Expression

Methylation

….

Measured via **microarrays** or **deep sequencing**

Tabular Data

Example Tasks:

Identify shared function of genes

- Similar regulation can indicate functional relationships

Compare regulation between conditions

- Differential regulation between conditions can indicate functional relevance

Identify subtypes

- Differential regulation in “one” disease can reveal subtypes

Scale of Tables

Need different approaches for “normal” and “high-dimensional” tables.

How many dimensions?

~50 – tractable with “just” vis

~1000 – need analytical methods

How many records?

~ 1000 – “just” vis is fine

>> 10,000 – need analytical methods

Homogeneity

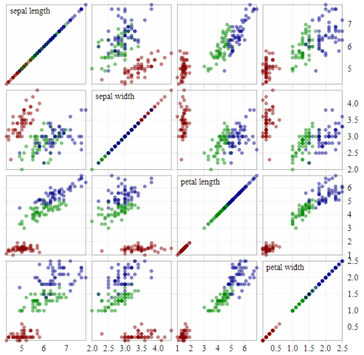
Same data type?

Same scales?

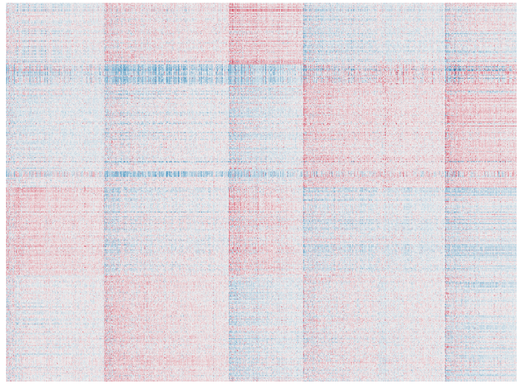
	EGFR Exp	EGFR CN	EGFR M
Bob	*1.45*	*1*	*Y*
Alice	*0.21*	*3*	*N*
Chris	*1.12*	*1*	*N*

	EGFR Ex	BRCA Ex	TNF Ex
Bob	*1.45*	*1.47*	*3.2*
Alice	*0.21*	*1.8*	*1.1*
Chris	*1.12*	*0.27*	*0.9*

Analytic Component

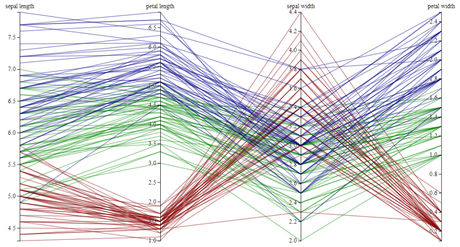


Scatterplot Matrices
[Bostock]

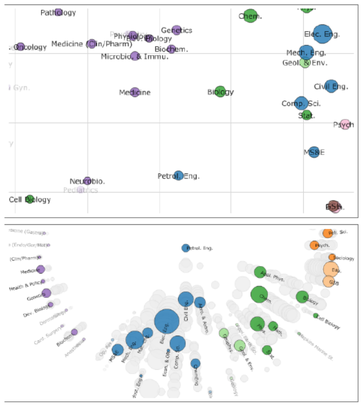


Pixel-based visualizations / heat maps

Multidimensional Scaling
[Doerk 2011]



Parallel Coordinates
[Bostock]



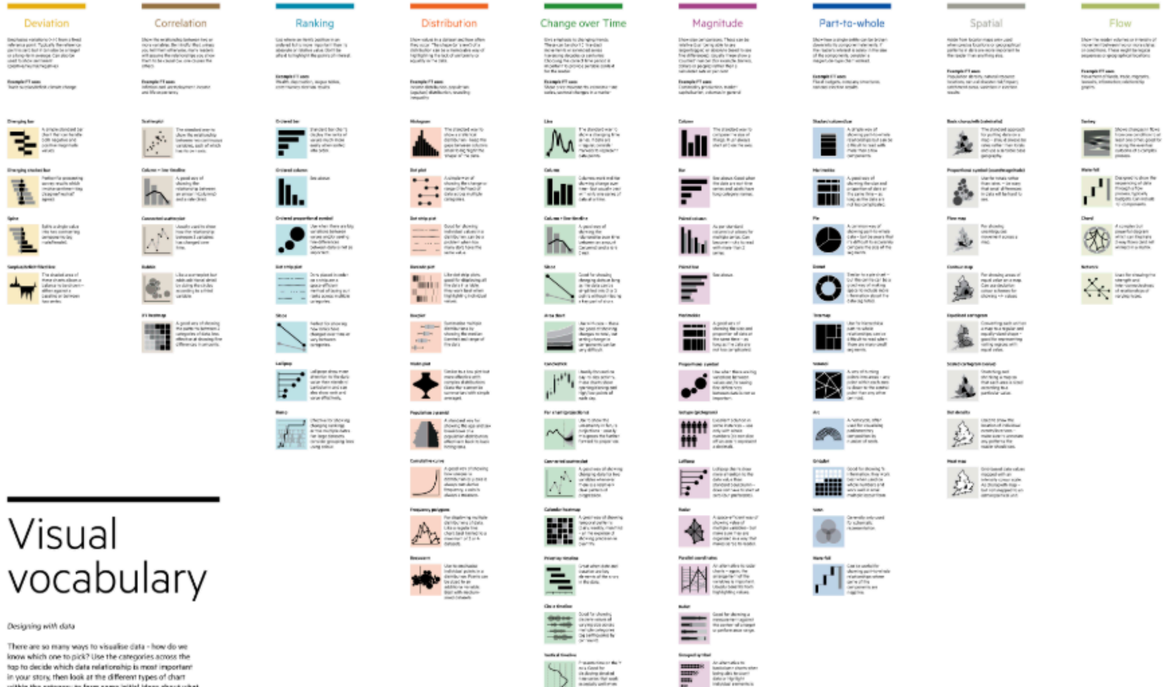
[Chuang 2012]

no / little analytics ⟷ **strong analytics component**

Techniques and Tasks

Magnitude Distribution Deviation Correlation

Ranking Part to whole Change over Time



https://github.com/ft-interactive/chart-doctor/tree/master/visual-vocabulary
https://gramener.github.io/visual-vocabulary-vega/#/Magnitude/

Aggregating Large Data Vectors

Instead of showing all data points, show a data's distribution

Pro: compact representation

Con: Works only if data is "well behaved" for the type of distribution visualization.

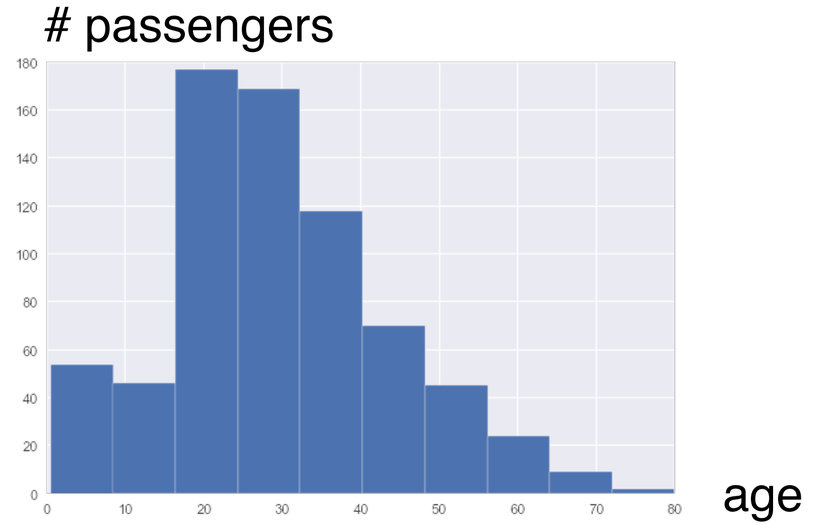
Histogram

Good #bins hard to predict

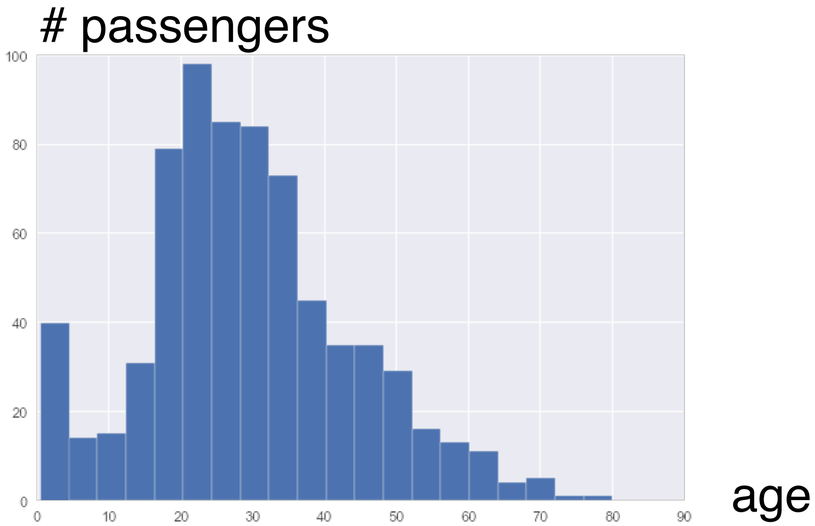
make interactive!

rules of thumb:

- #bins = sqrt(n)
- #bins = log2(n) + 1

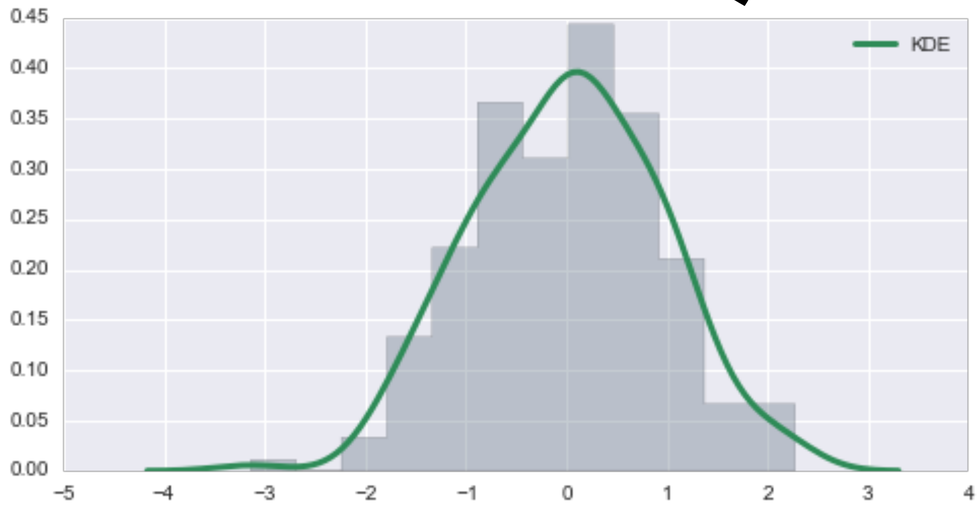


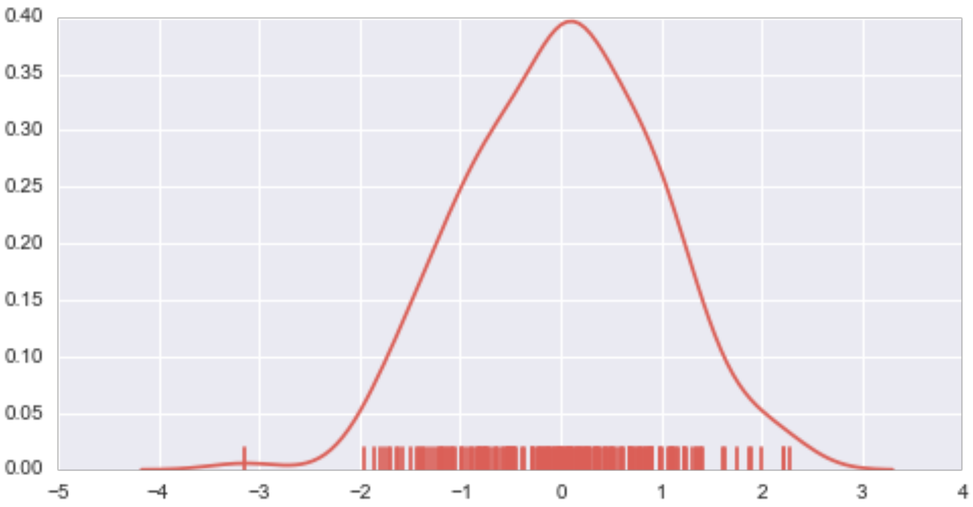
10 Bins

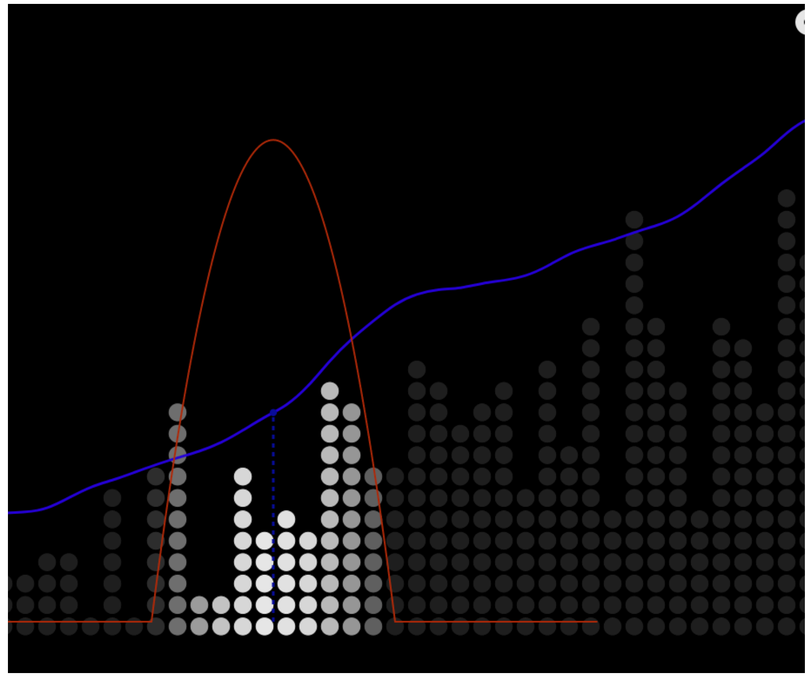


20 Bins

Density Plots (Kernel Density Estimation)







http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

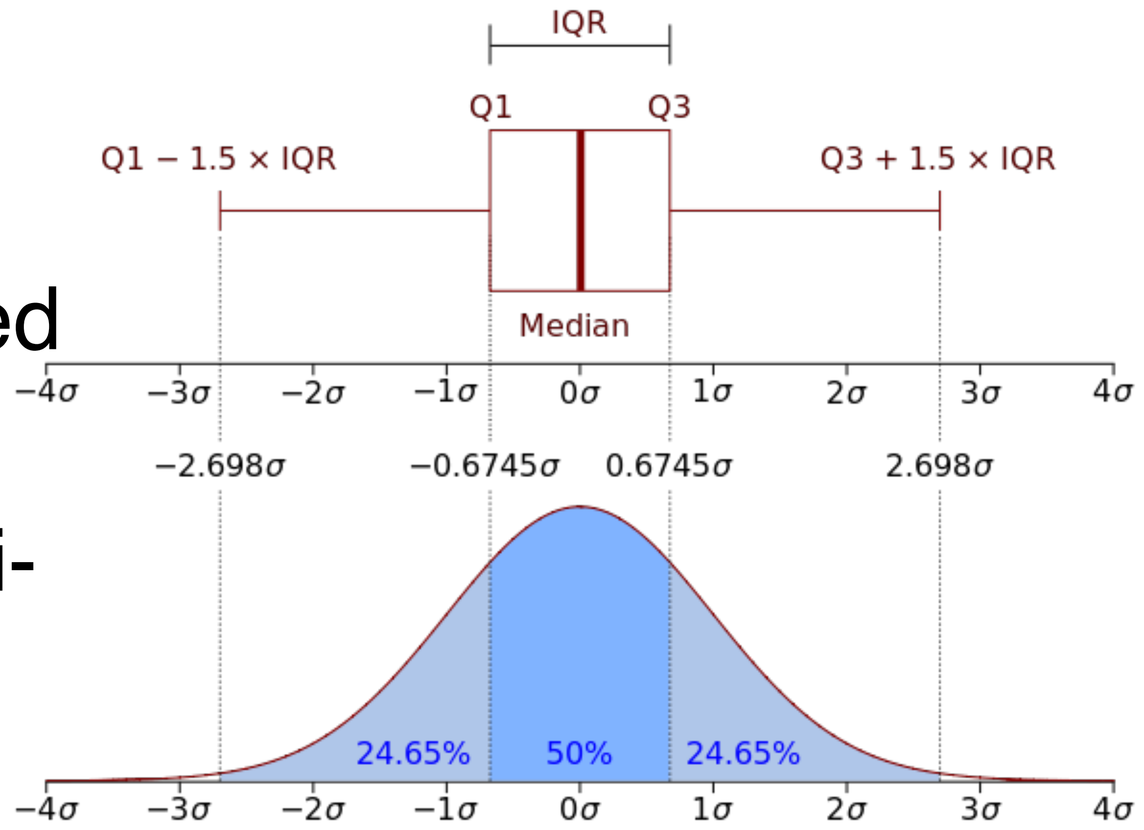
Box Plots

aka Box-and-Whisker Plot

Show outliers as points!

Bad for non-normal distributed data

Especially bad for bi- or multi-modal distributions



Wikipedia

One Boxplot, Four Distributions

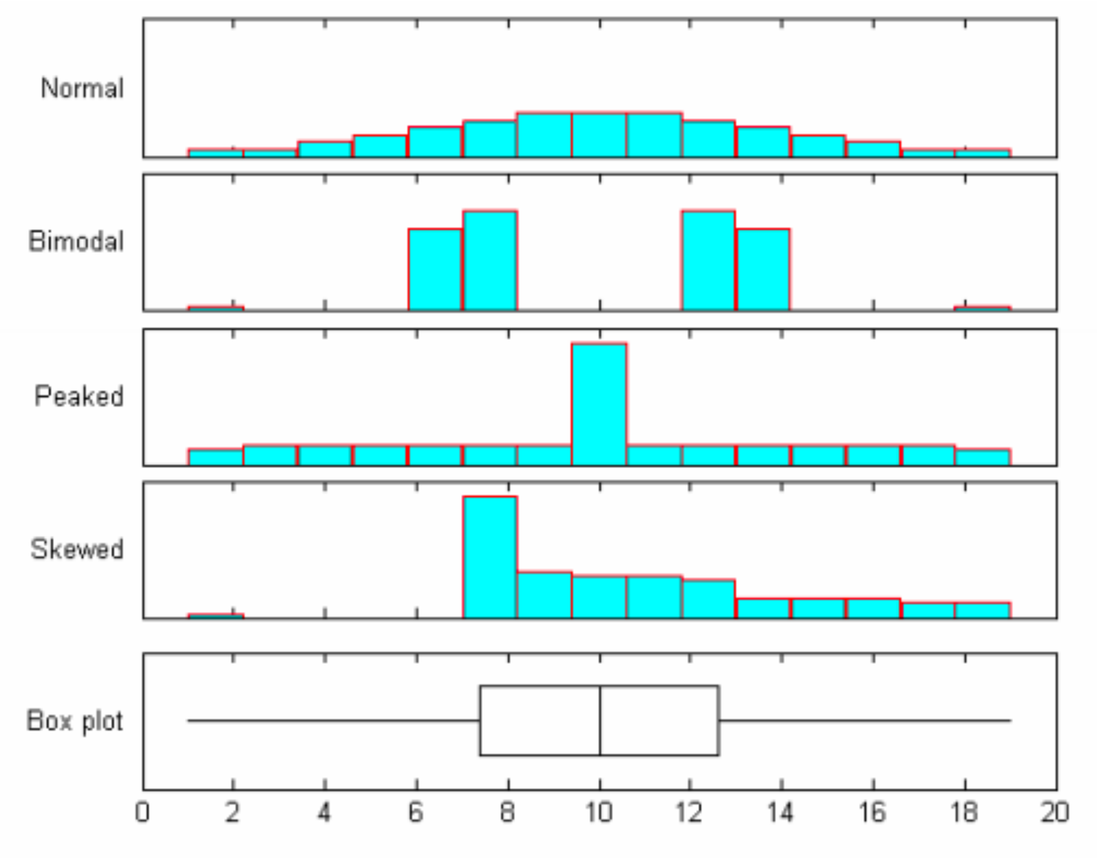
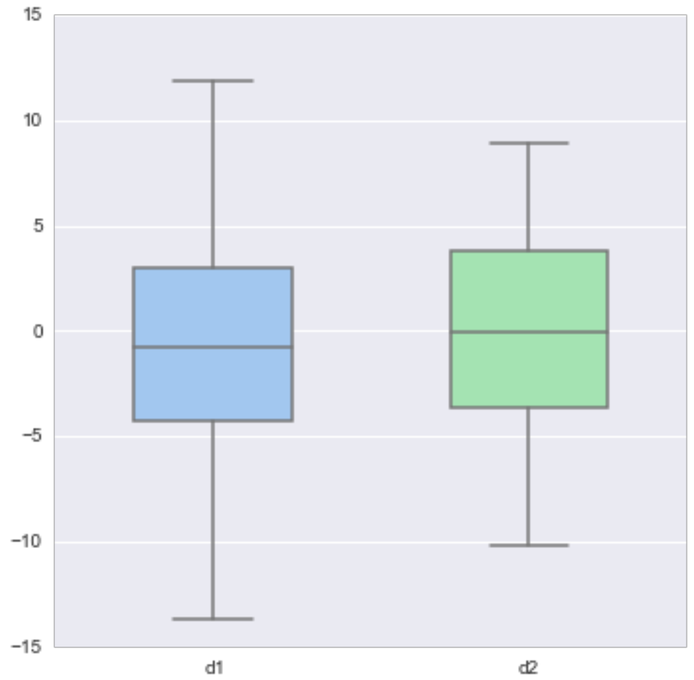


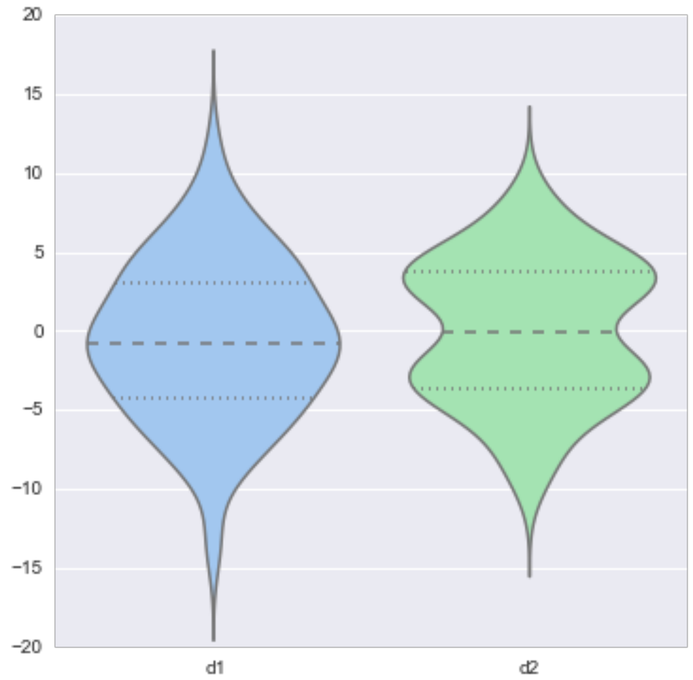
Figure 1: Histograms and box plot: four samples each of size 100

http://stat.mq.edu.au/wp-content/uploads/2014/05/Can_the_Box_Plot_be_Improved.pdf

Violin Plot

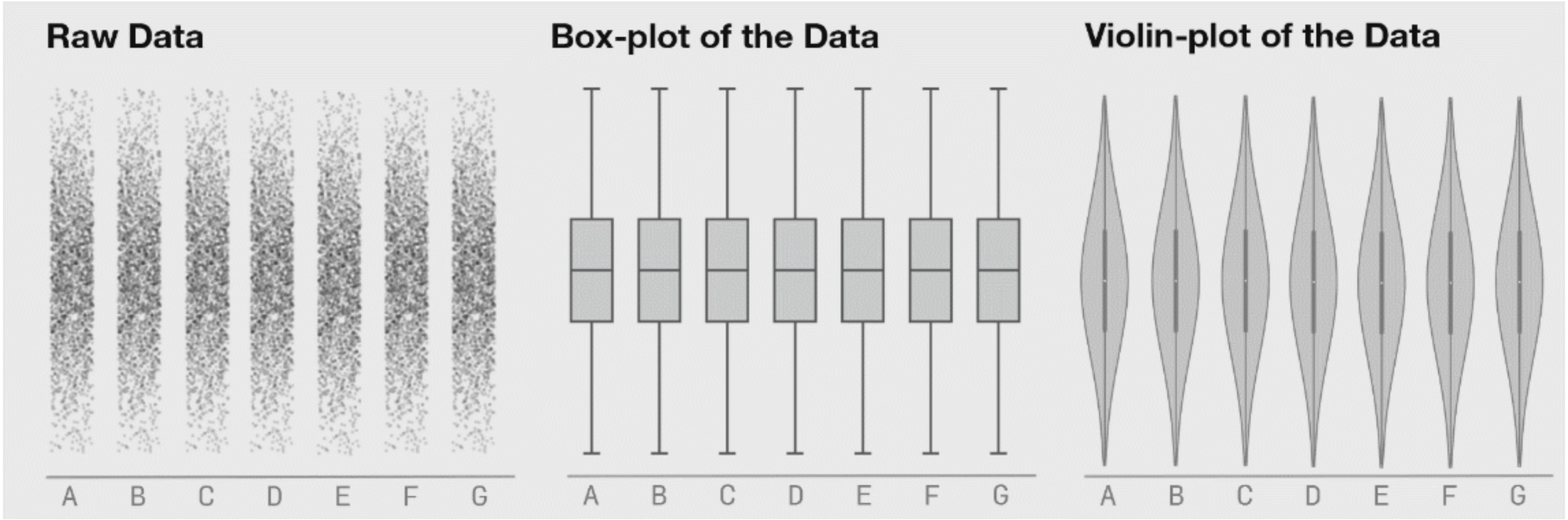
= Box Plot + Probability Density Function





http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

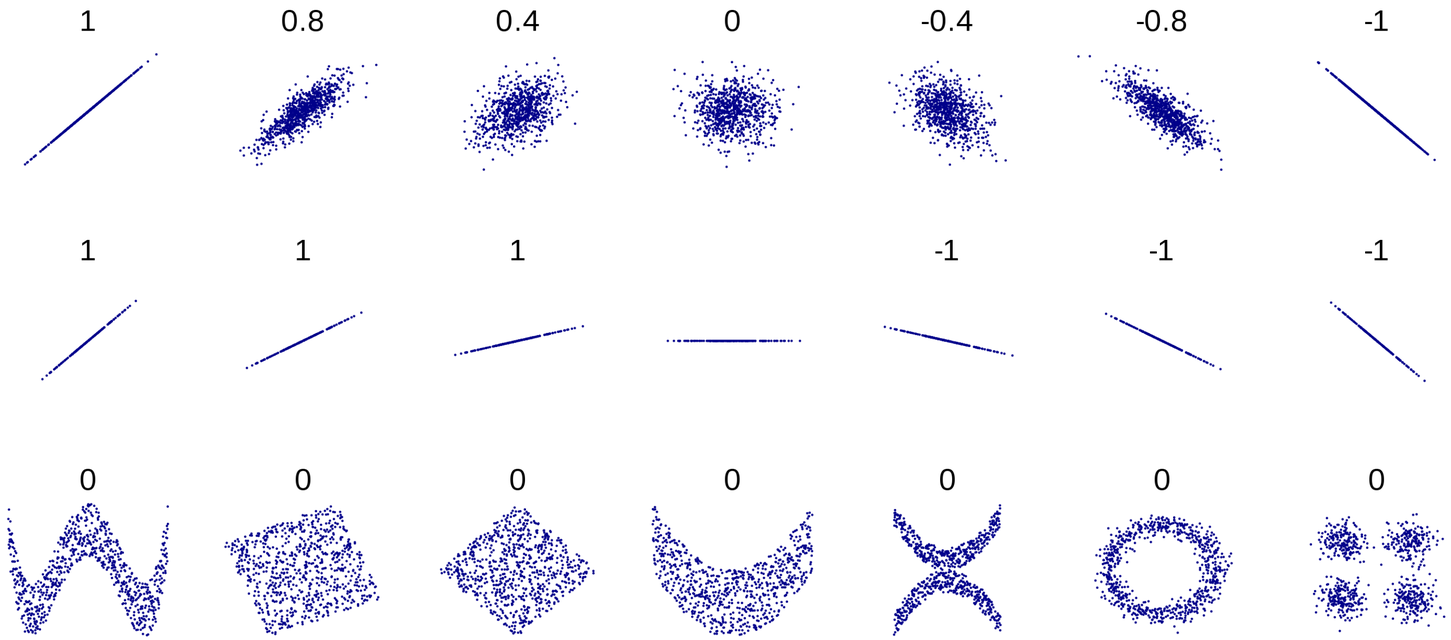
Different Distributions



https://blog.bioturing.com/2018/05/16/5-reasons-you-should-use-a-violin-graph/

Correlation

How do two or more variables behave relative to each other?



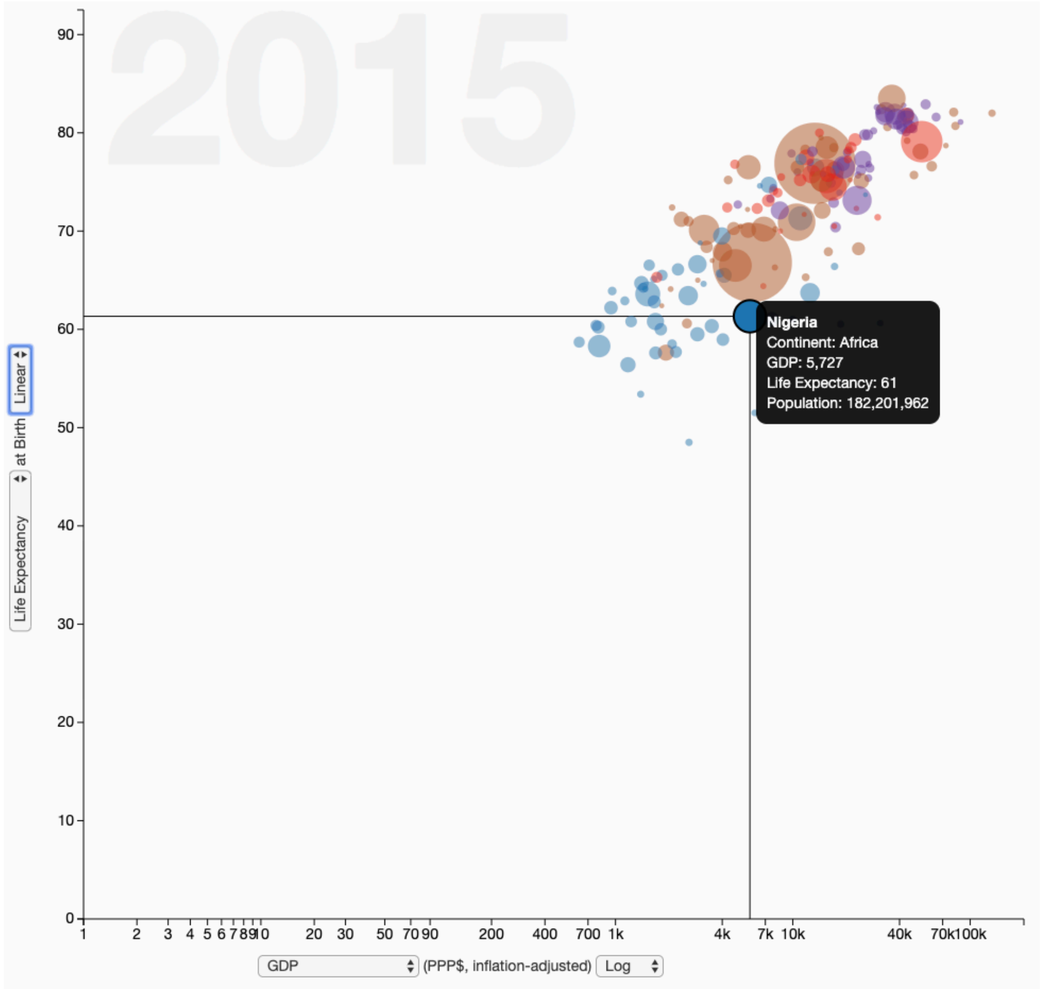
By DenisBoigelot, CC0, https://commons.wikimedia.org/w/index.php?curid=15165296

Scatterplots

Two orthogonal axis visualizing one dimension each.

How to encode the mark?

How to deal with many points?

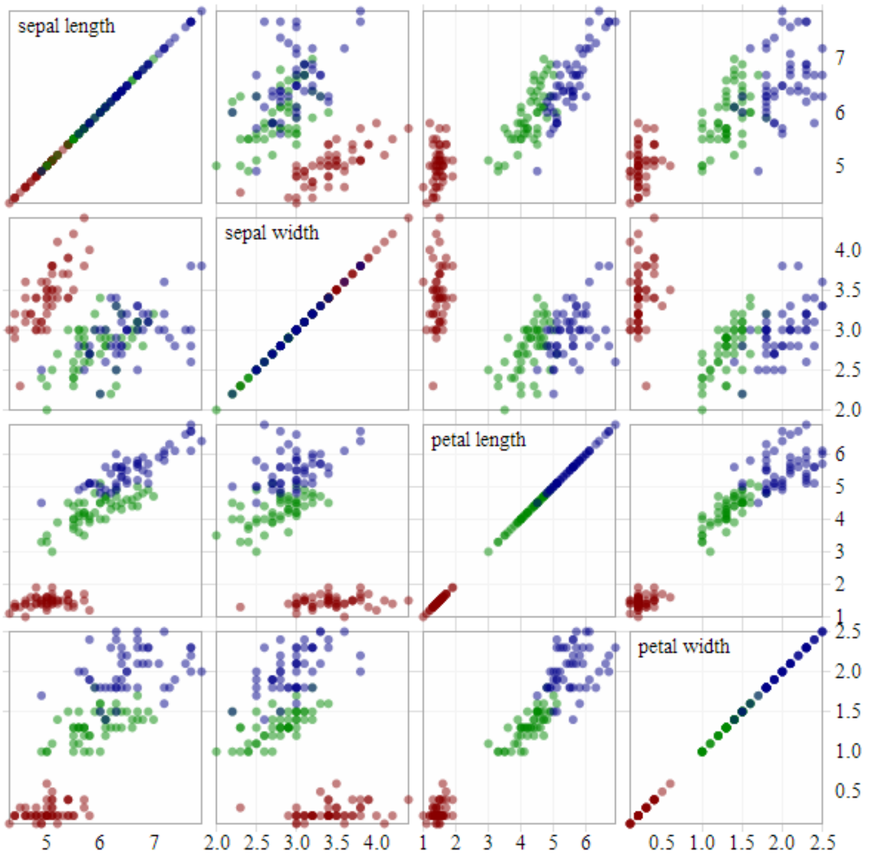


Scatterplot Matrices (SPLOM)

Matrix of size d*d

Each row/column is one dimension

Each cell plots a scatterplot of two dimensions

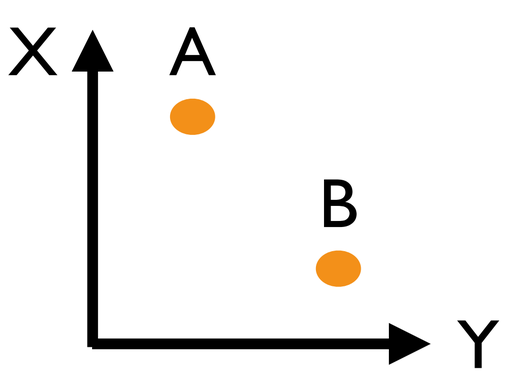


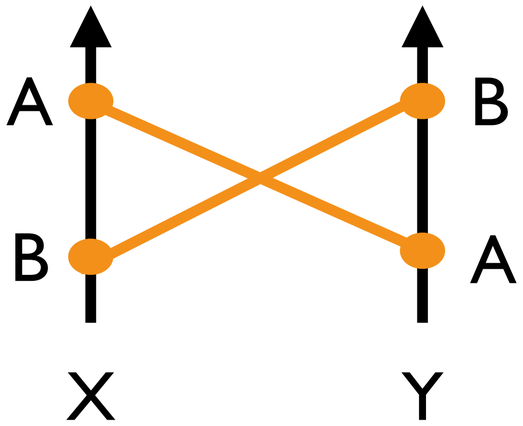
Parallel Coordinates (PC)

Inselberg 1985

Axes represent attributes

Lines connecting axes represent items





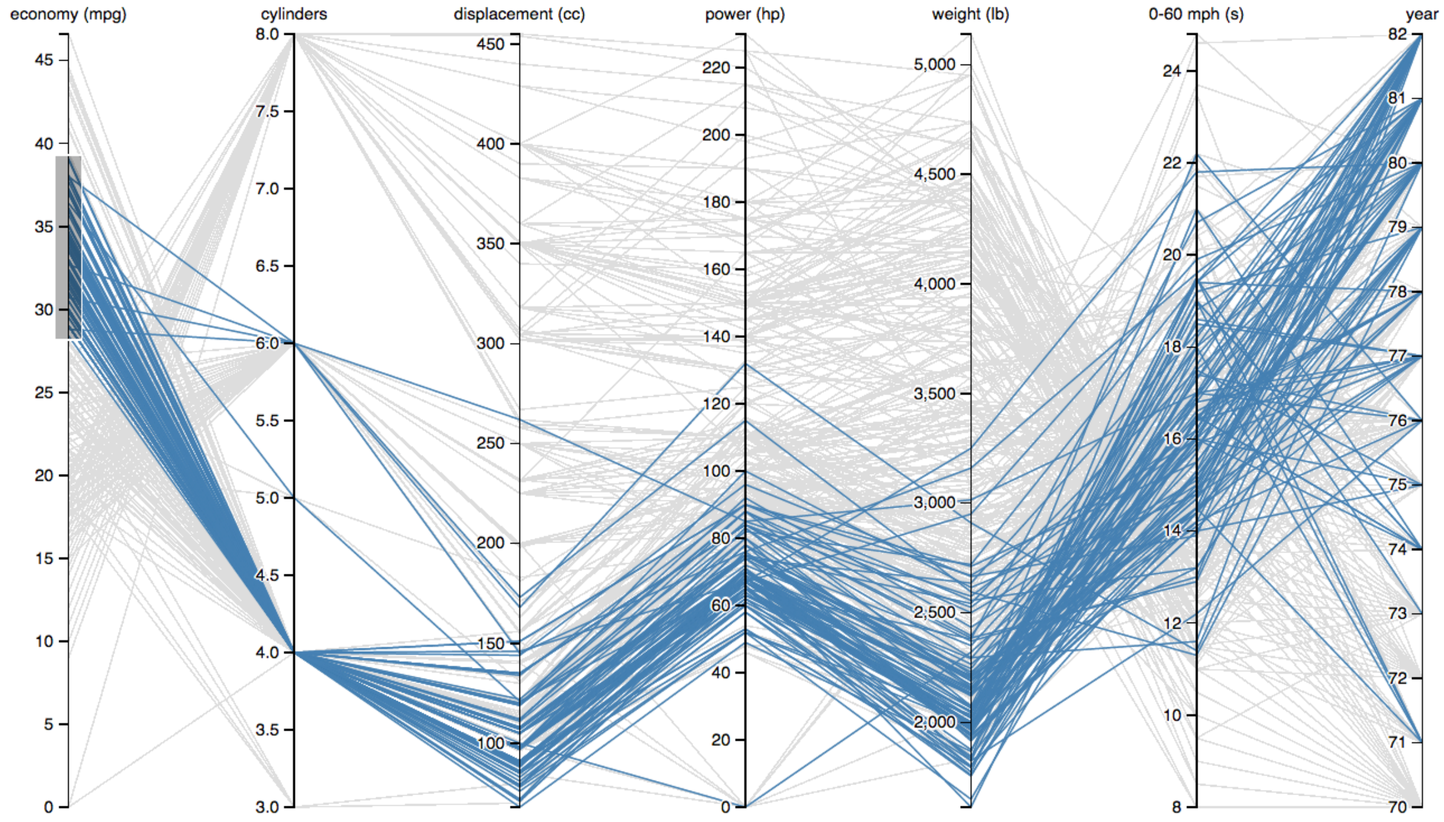
Parallel Coordinates

Each axis represents dimension

Lines connecting axis represent records

Suitable for

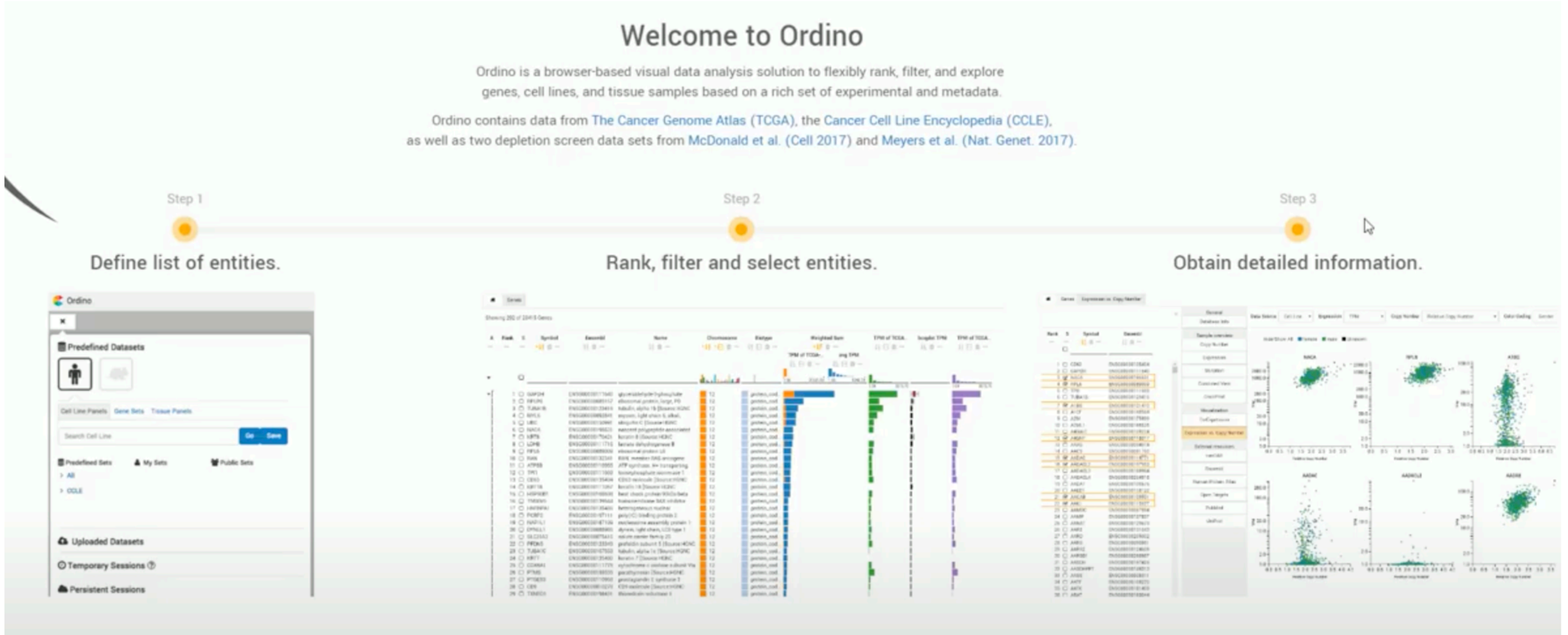
- all tabular data types
- heterogeneous data



Tabular Representation

Like spreadsheet: each variable in it’s own column

Visual encodings to make it scalable



Pixel Based Displays

Each cell is a “pixel”, value encoded in color / value

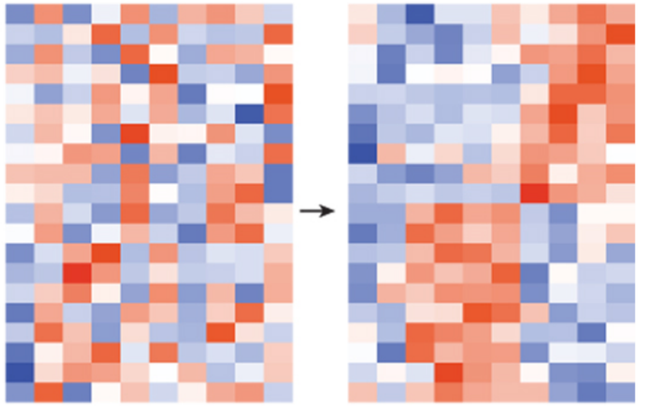
Ordering critical for interpretation

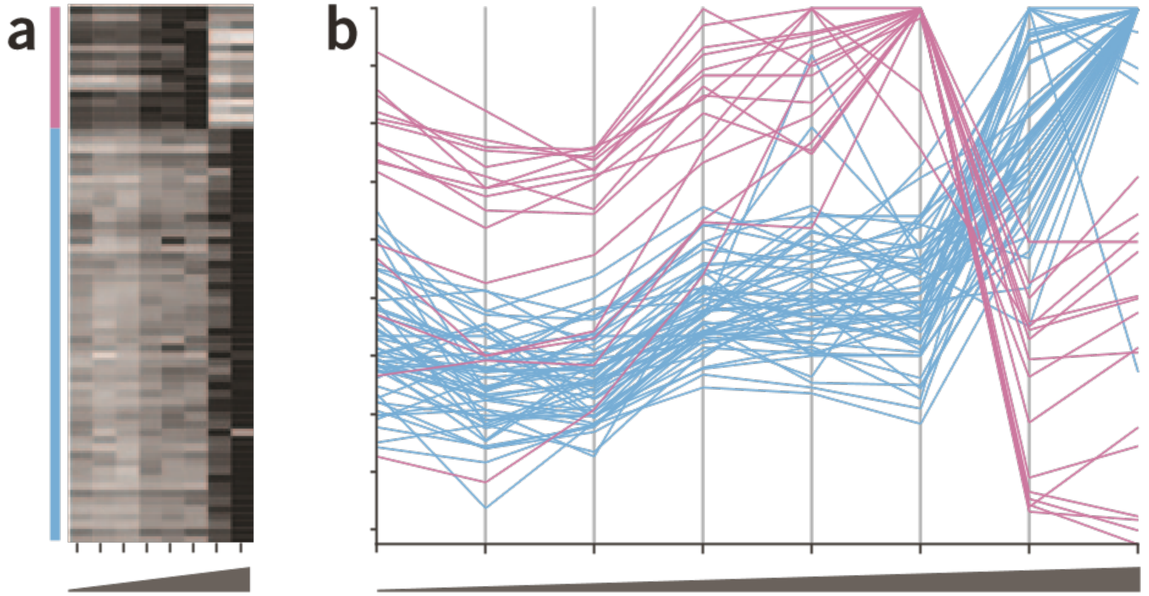
If no ordering inherent, clustering is used

Scalable – 1 px per item

Good for homogeneous data

same scale & type





[Gehlenborg & Wong 2012]

Clustering

Classification of items into “similar” bins

Based on similarity measures

- Euclidean distance, Pearson correlation, ...

Partitional Algorithms

- divide data into set of bins
- # bins either manually set (e.g., k-means) or automatically determined (e.g., affinity propagation)

Hierarchical Algorithms

- Produce “similarity tree” – dendrogram

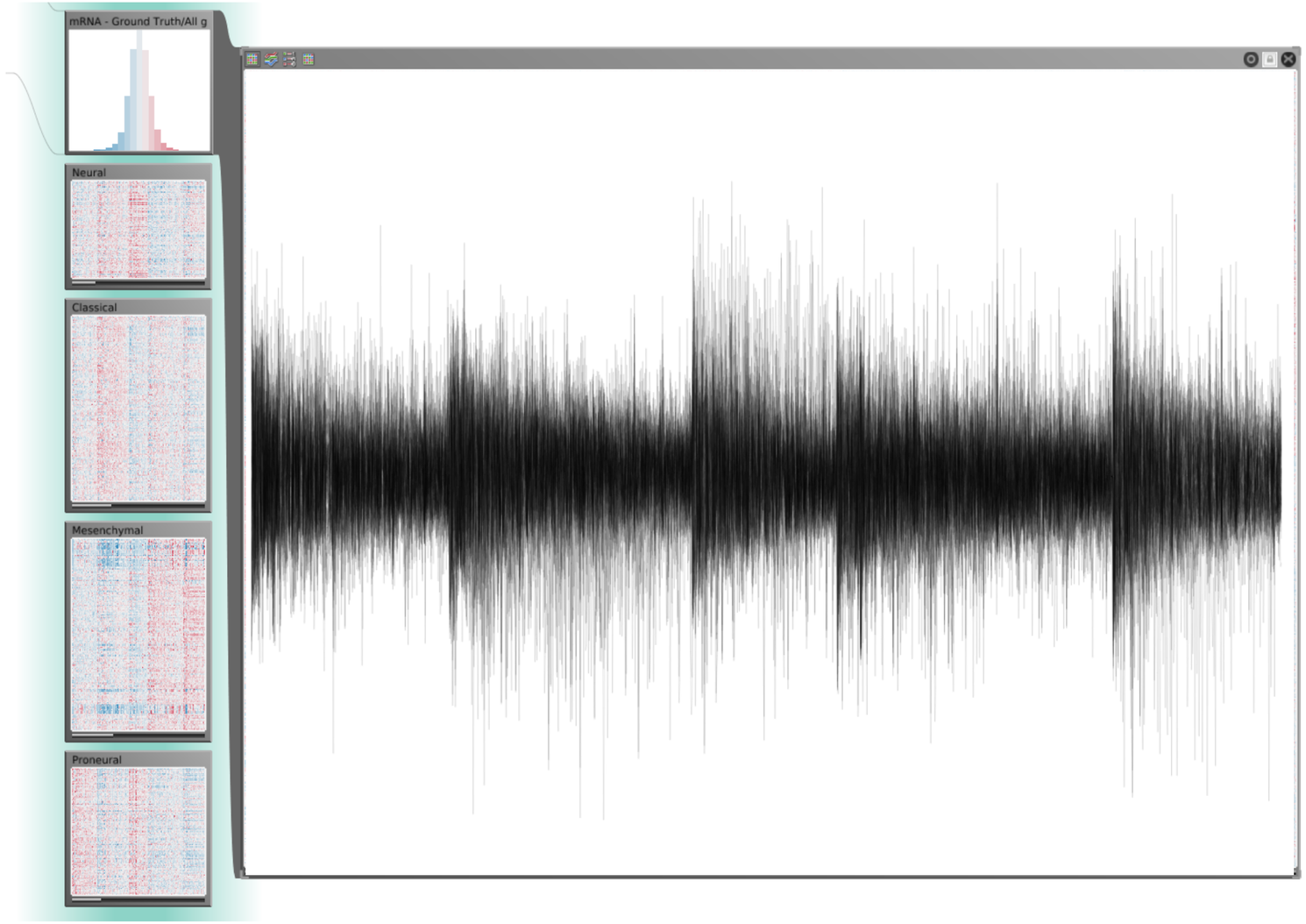
Bi-Clustering

- Clusters dimensions & records

Fuzzy clustering

- allows occurrence of elements in multiples clusters

Clustered Heat Map



Hierarchical Clustering

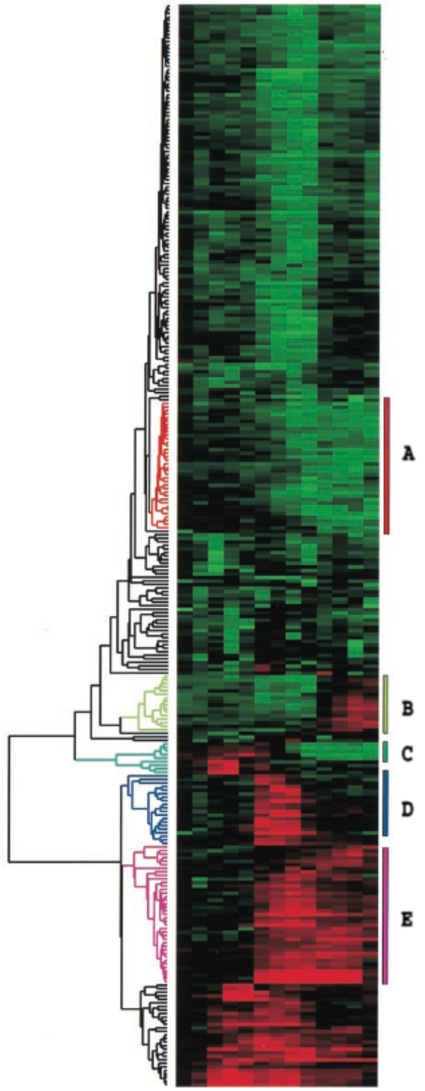
Two types:

agglomerative clustering

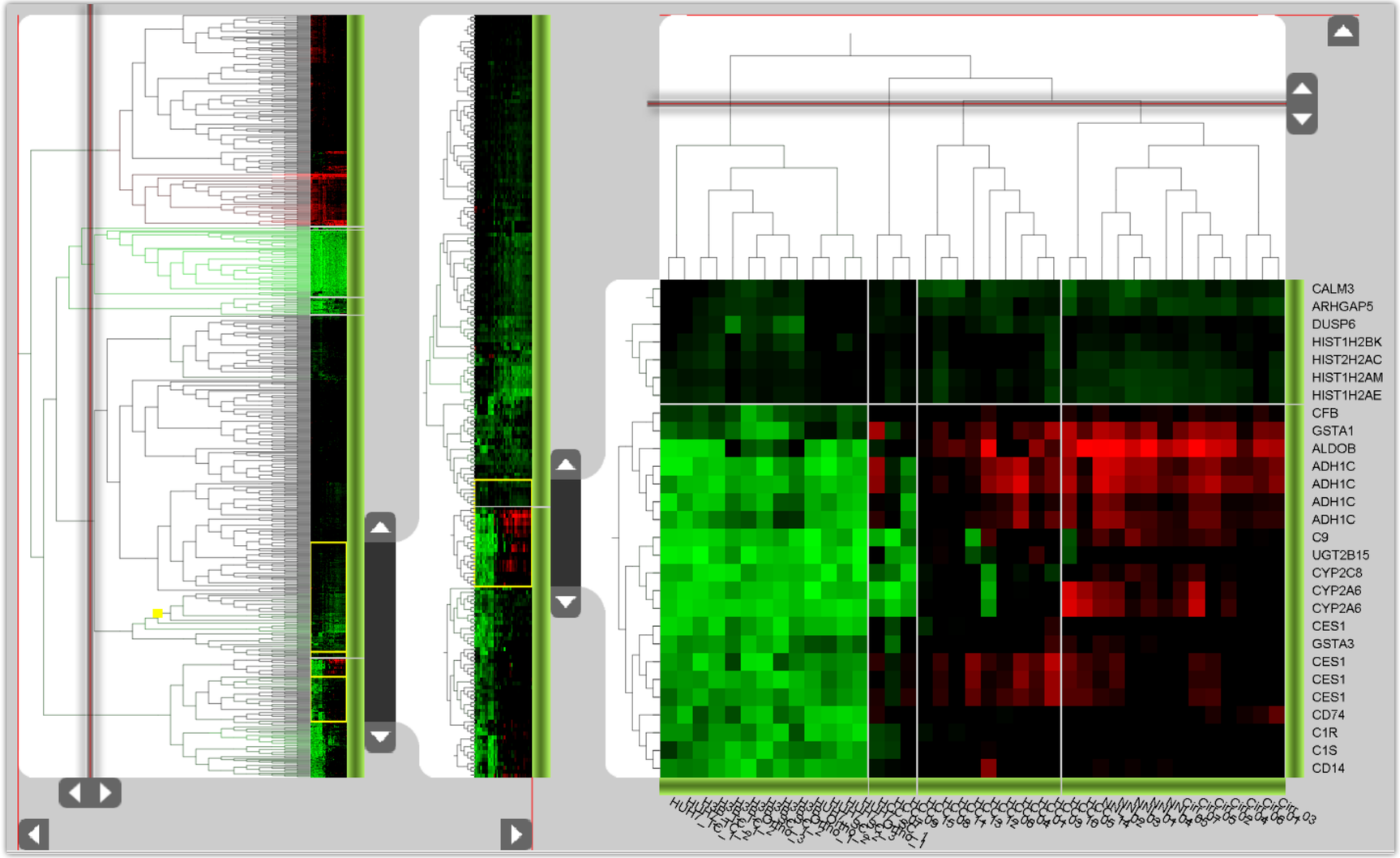
start with each node as a cluster and merge

divisive clustering

start with one cluster, and split

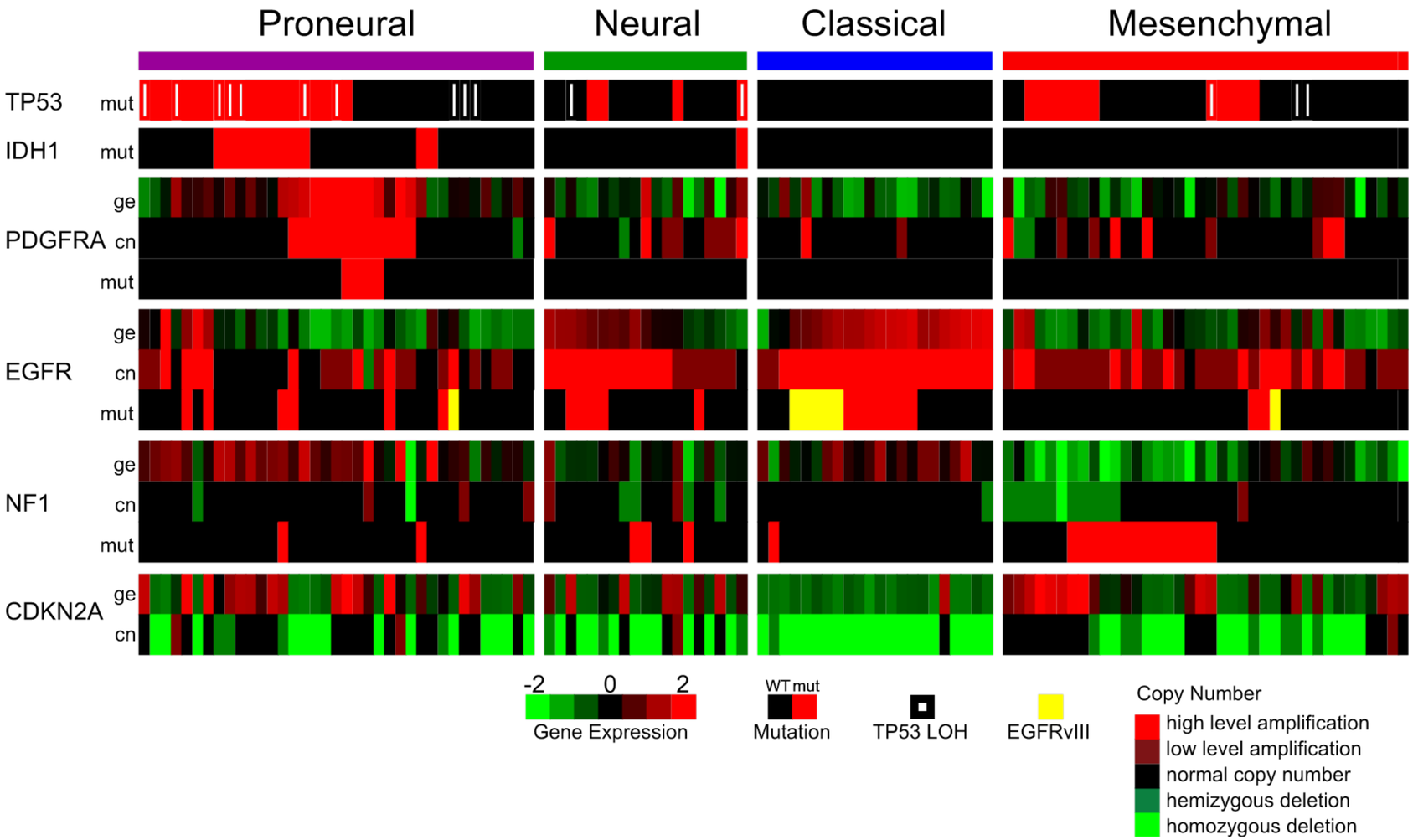


F+C Approach, with Dendrograms



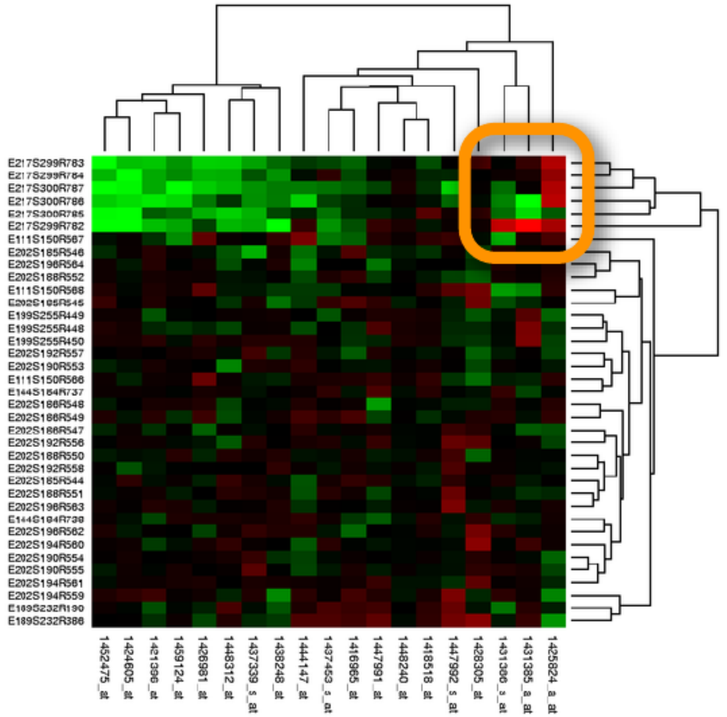
[Lex, PacificVis 2010]

Heterogeneous Data?

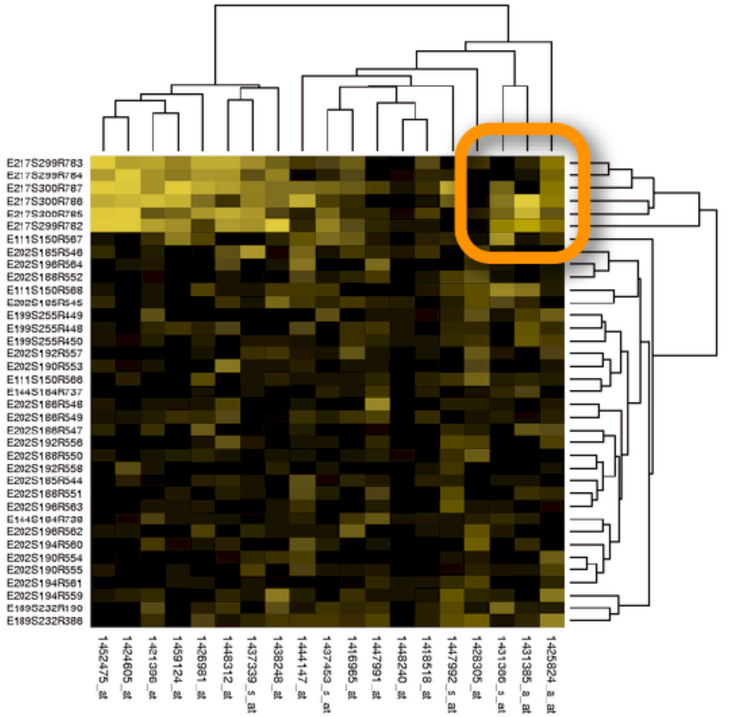


[Verhaak 2012]

Bad Color Mapping

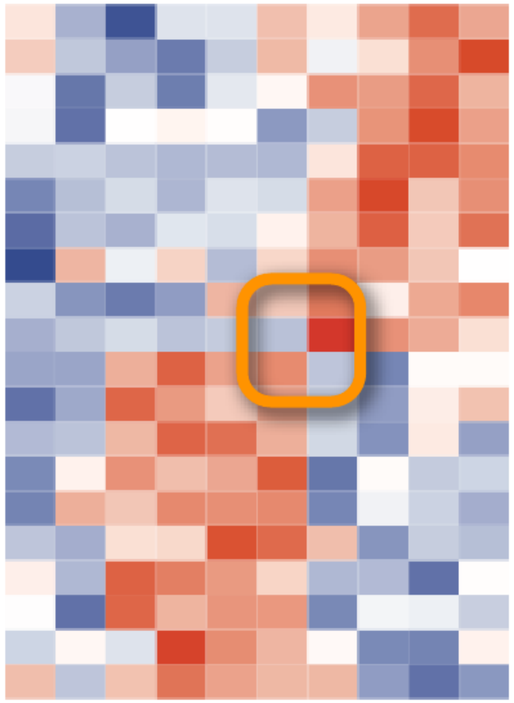


Normal Vision

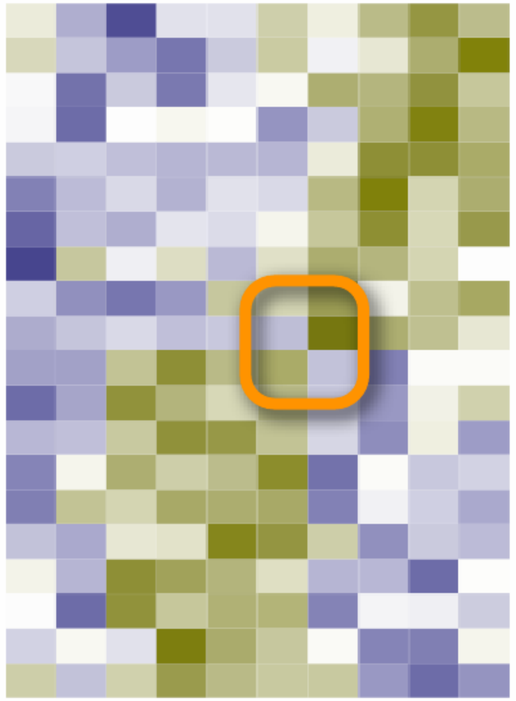


Deuteranope Vision
("Red-Green Blindness")

Good Color Mapping

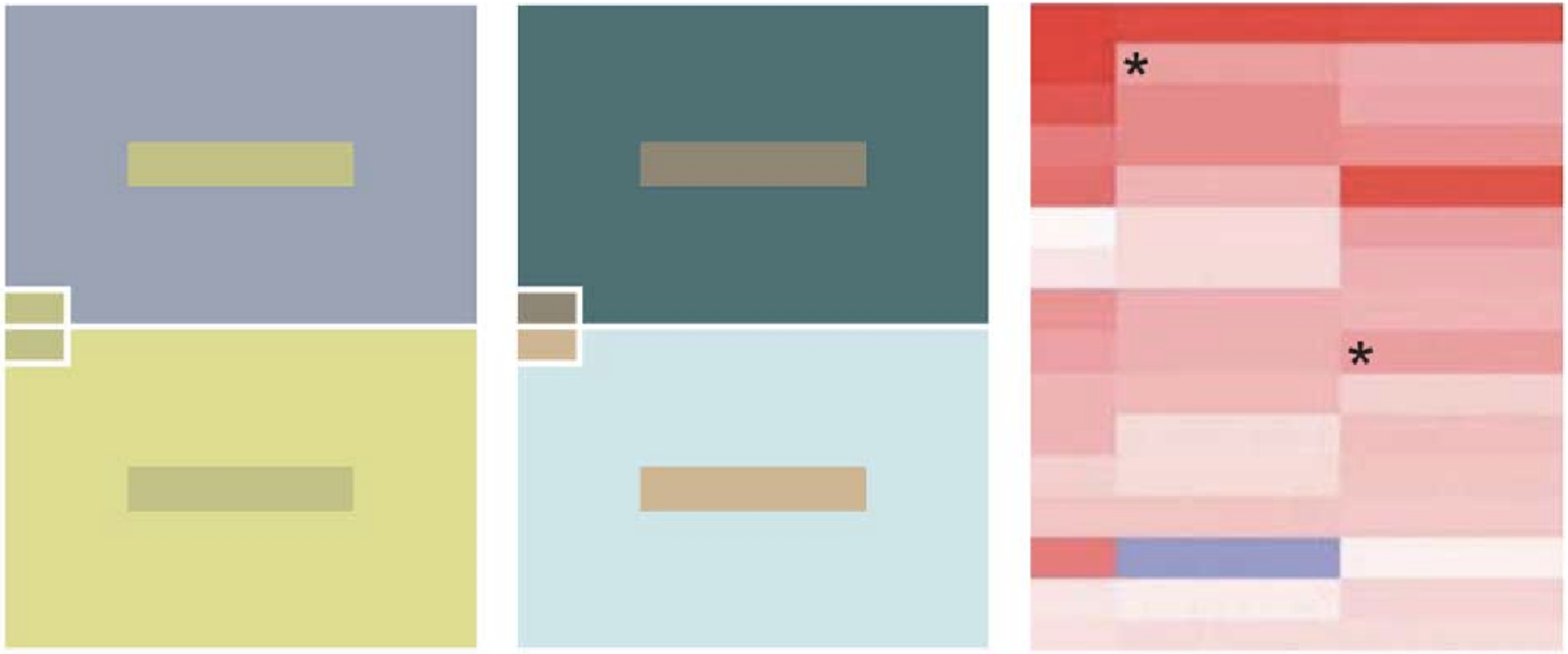


Normal Vision

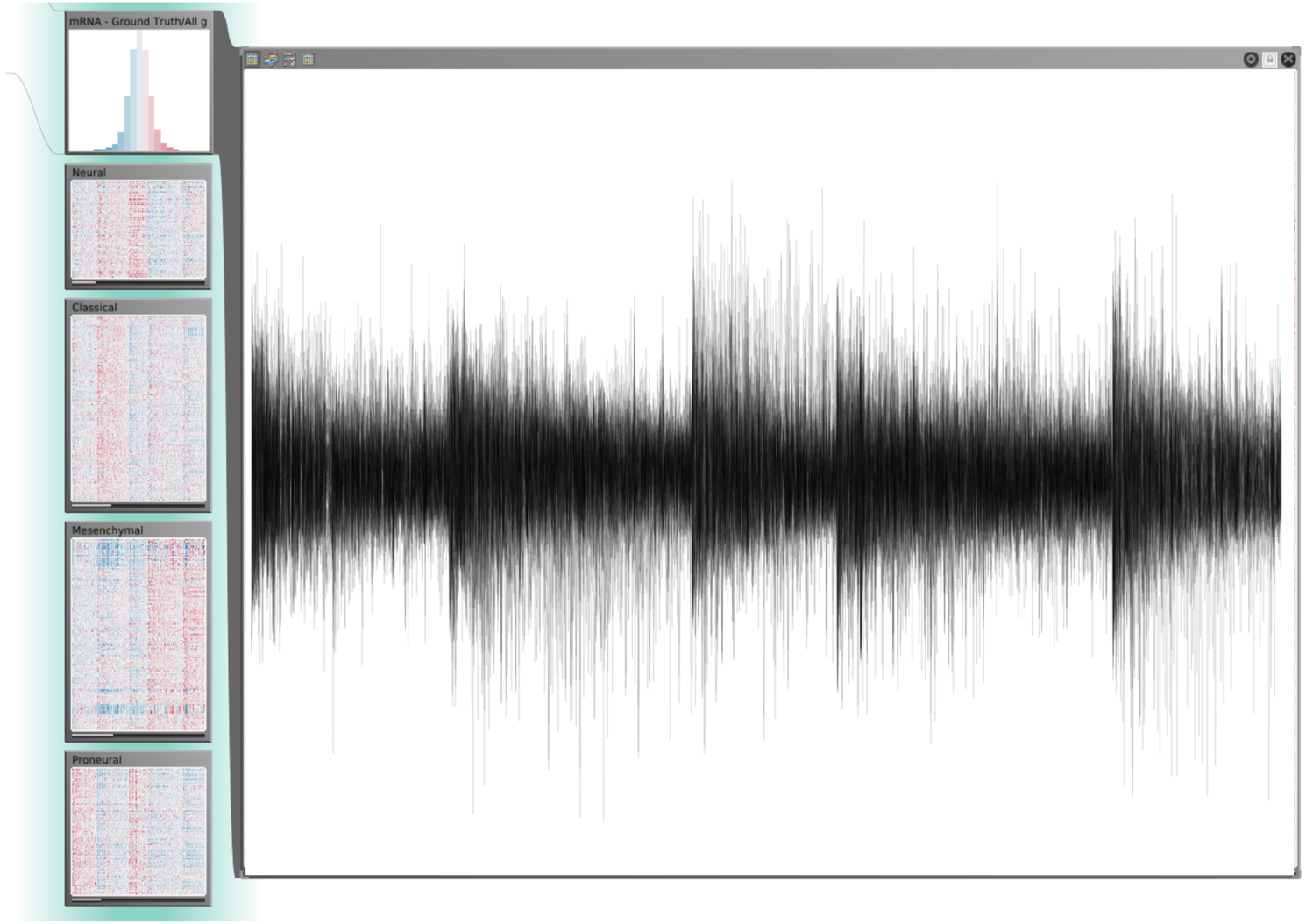


Deuteranope Vision
("Red-Green Blindness")

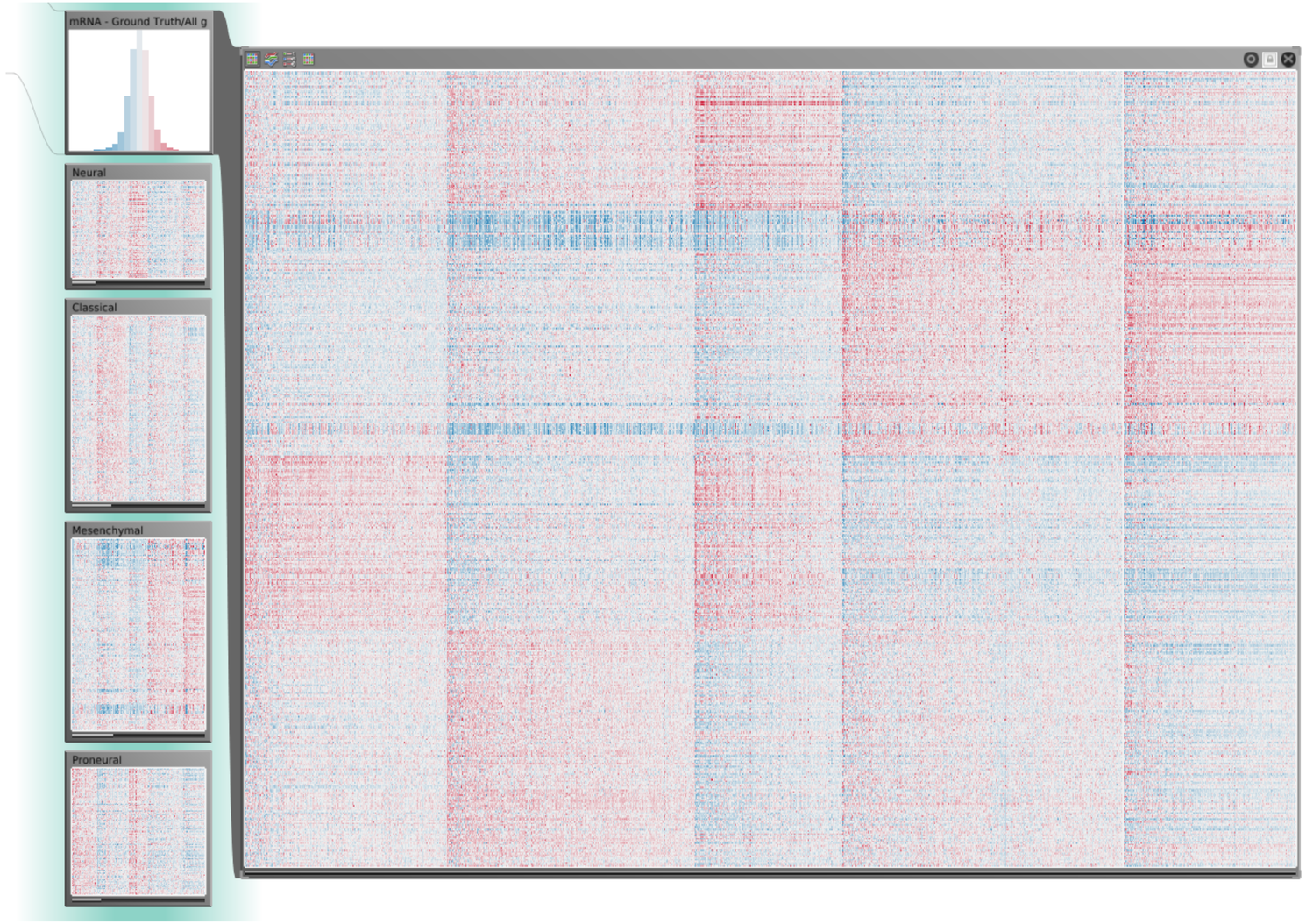
Color is relative!



Clustered Heat Map



Example: Cancer Subtypes



Dimensionality Reduction

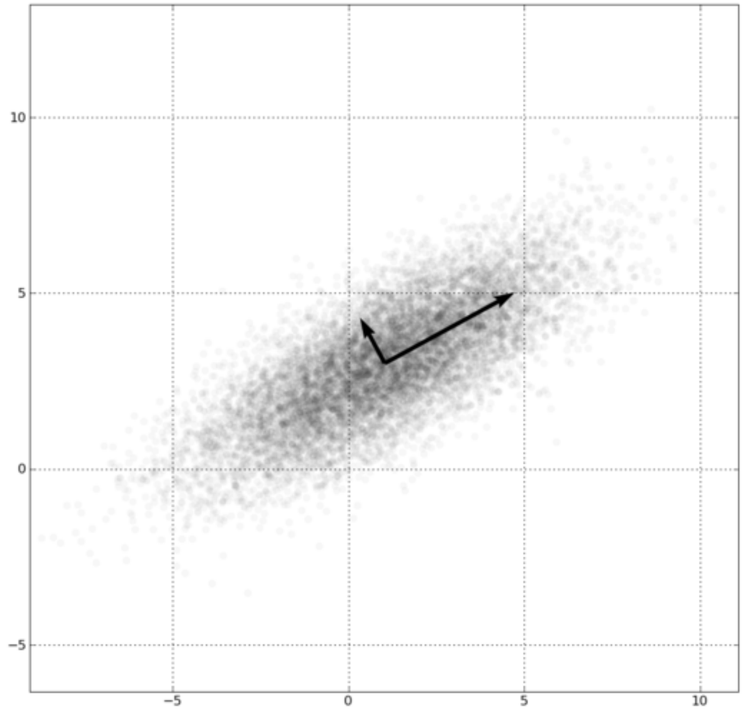
Reduce high dimensional to lower dimensional space

Preserve as much of variation as possible

Plot lower dimensional space

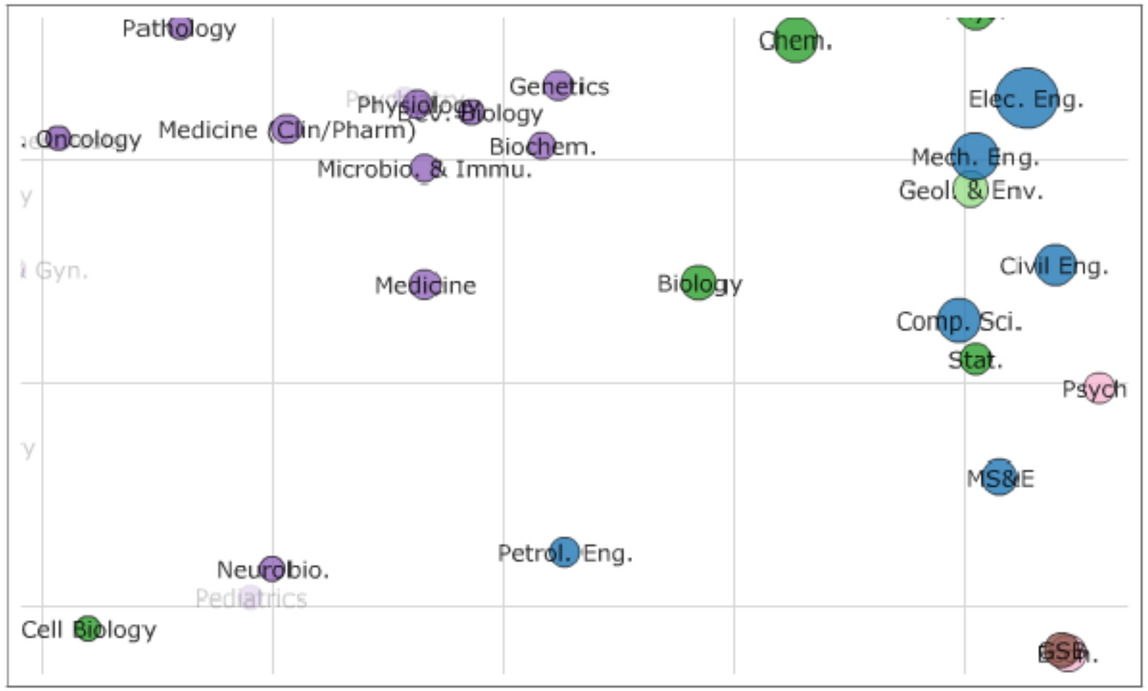
Principal Component Analysis (PCA)

linear mapping, by order of variance

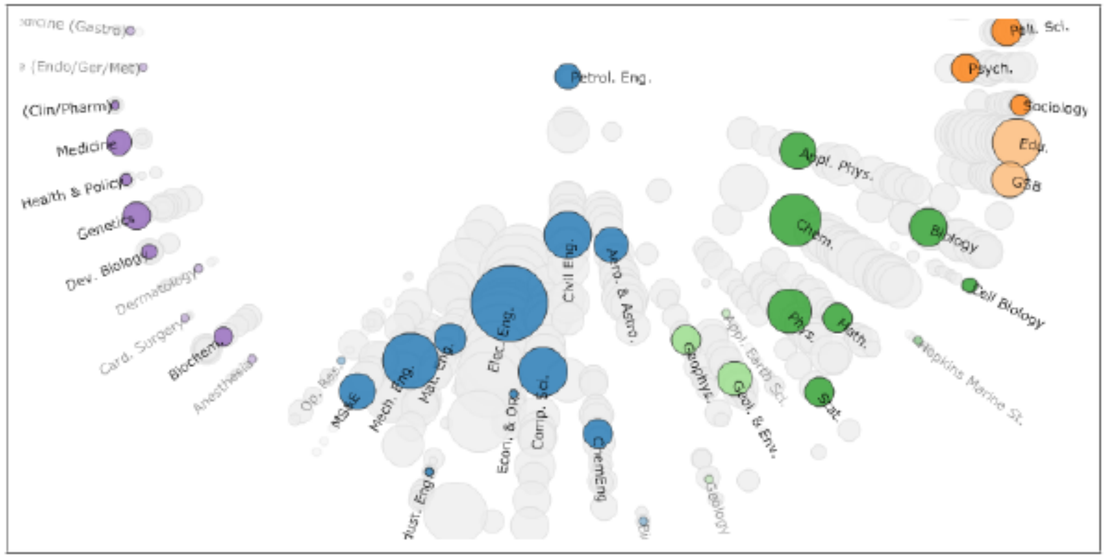


Can we Trust Dimensionality Reduction?

Topical distances between departments in a 2D projection



Topical distances between the selected Petroleum Engineering and the others.



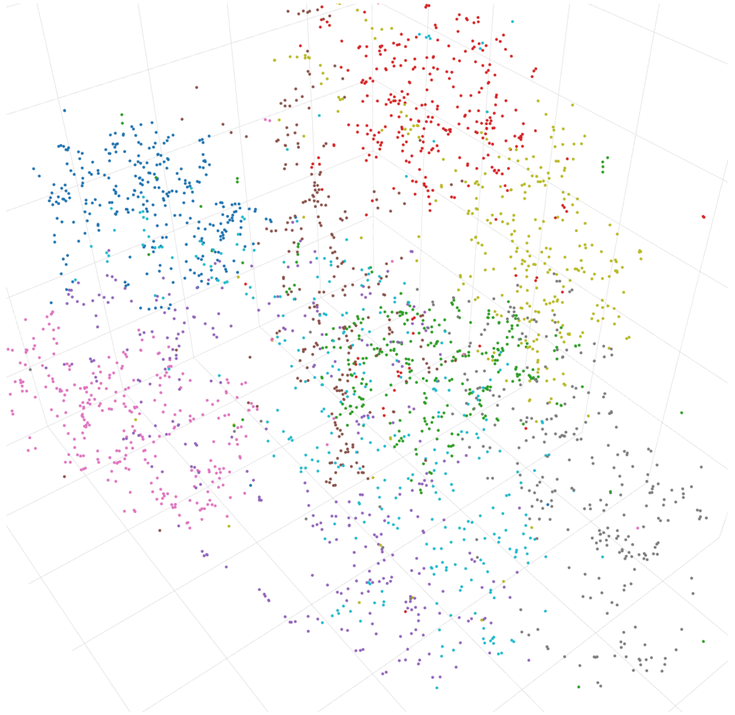
[Chuang et al., 2012]

http://www-nlp.stanford.edu/projects/dissertations/browser.html

t-SNE

t-distributed stochastic neighbor embedding

non-linear algorithm: different transformations for different regions



Visualizing data using t-SNE, Maaten and Hinton, 2008

How to Use t-SNE Effectively

Although extremely useful for visualizing high-dimensional data, t-SNE plots can sometimes be mysterious or misleading. By exploring how it behaves in simple cases, we can learn to use it more effectively.

MARTIN WATTENBERG
Google Brain

FERNANDA VIÉGAS
Google Brain

IAN JOHNSON
Google Cloud

Oct. 13
2016

Citation:
Wattenberg, et al., 2016

Understanding UMAP

Andy Coenen, Adam Pearce | *Google PAIR*

Dimensionality reduction is a powerful tool for machine learning practitioners to visualize and understand large, high dimensional datasets. One of the most widely used techniques for visualization is t-SNE, but its performance suffers with large datasets and using it correctly can be challenging.

UMAP is a new technique by McInnes et al. that offers a number of advantages over t-SNE, most notably increased speed and better preservation of the data's global structure. In this article, we'll take a look at the theory behind UMAP in order to better understand how the algorithm works, how to use it effectively, and how its performance compares with t-SNE.

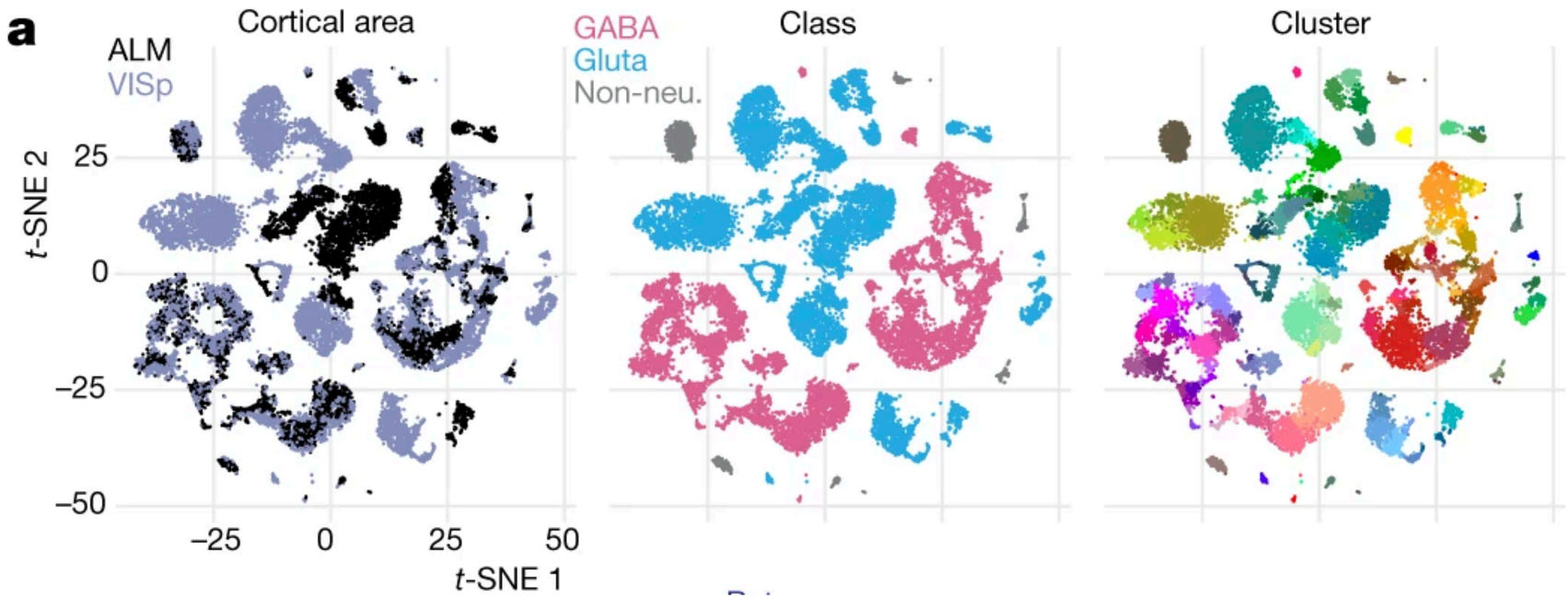
Figure 1: *Apply UMAP projection to various toy datasets, powered by umap-js.*

So what does UMAP bring to the table? Most importantly, UMAP is fast, scaling well in terms of

DR Example

Fig. 2: Comparison of gene expression differences among types across cortical areas.

From: Shared and distinct transcriptomic cell types across neocortical areas



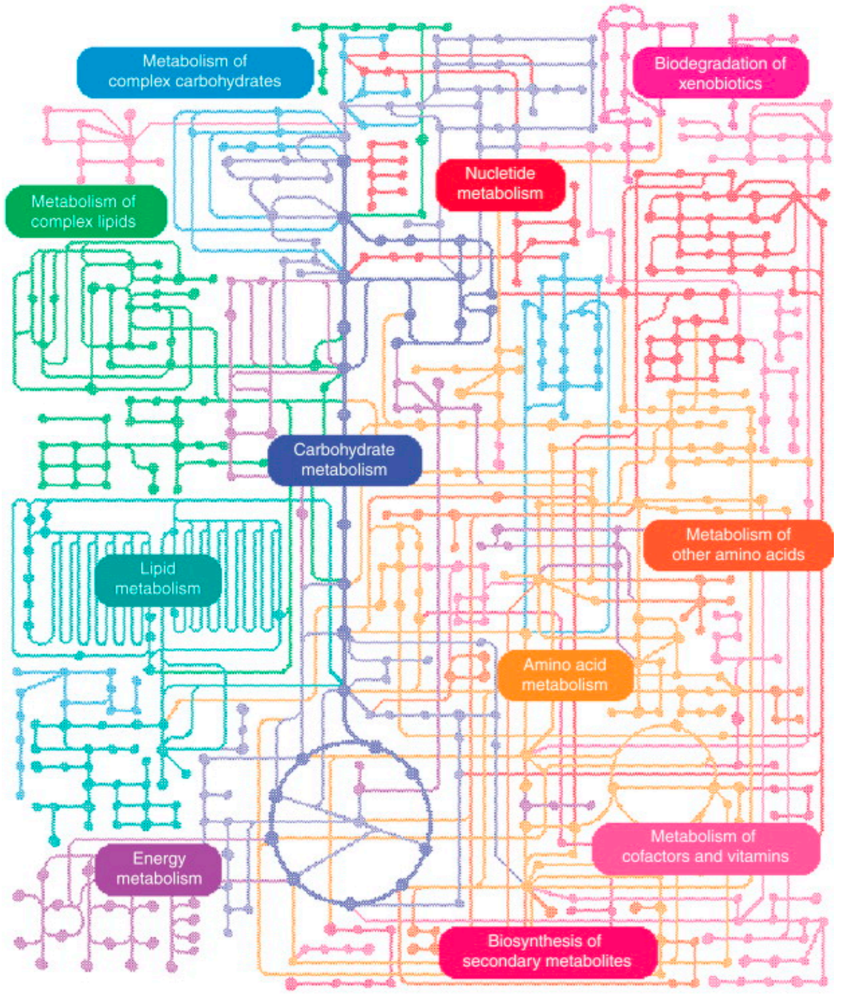
Tasic, B., Yao, Z., Graybuck, L.T. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature*

Systems Biology – Networks

How do things interact with each other?

- Metabolic processes: chemical reactions
  - E.g., food to energy, food to building blocks
- Signalling processes:
  - Cells respond to signals, coordinating function

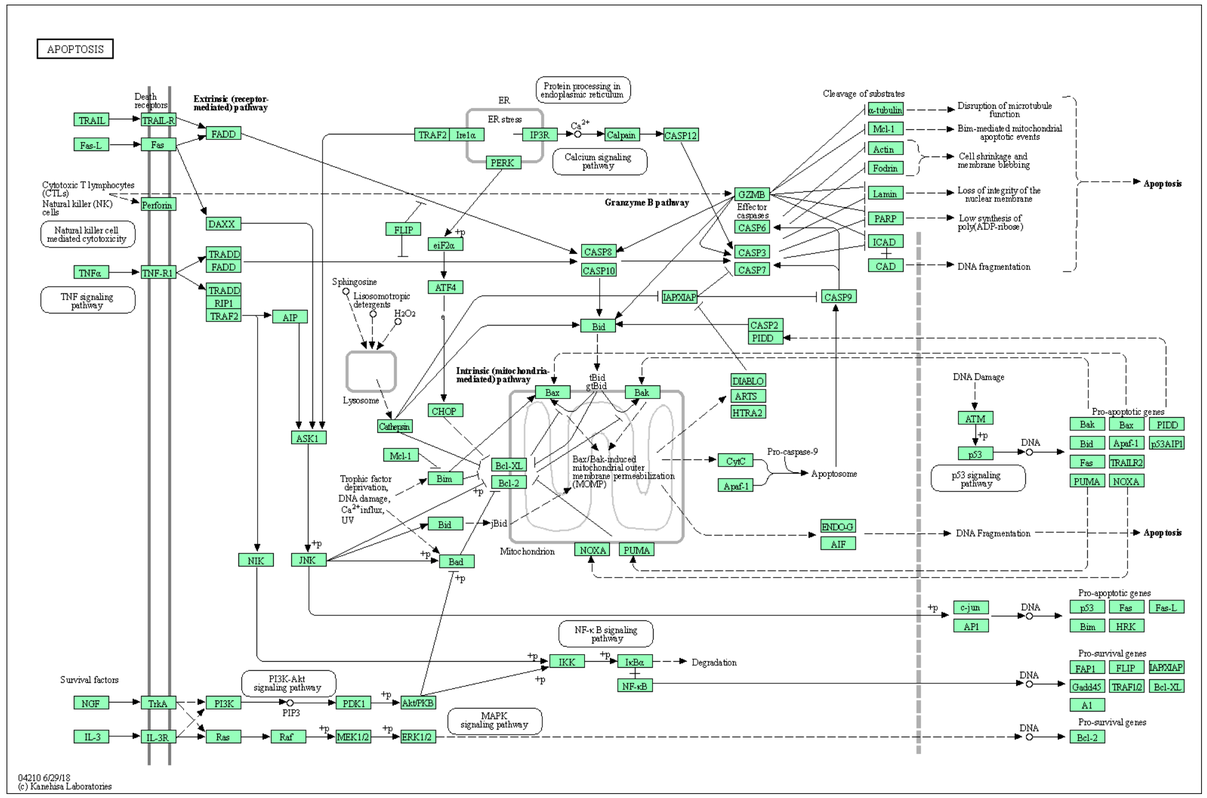
Understanding Omics Data in Context
Important to Understand Consequences on Processes



Metabolic and Signaling Networks

Curated Networks, available in various databased.

- KEGG
- Reactome
- Pathway Commons
- Biocarta…



Networks and Attributes

Attributes can influence topology

Path can be slow / blocked

- best route when driving depends on traffic
- biological network depends on many factors

Challenge: Data Scale & Heterogeneity

Large number of values

- Large datasets have more than 500 experiments

Multiple groups/conditions

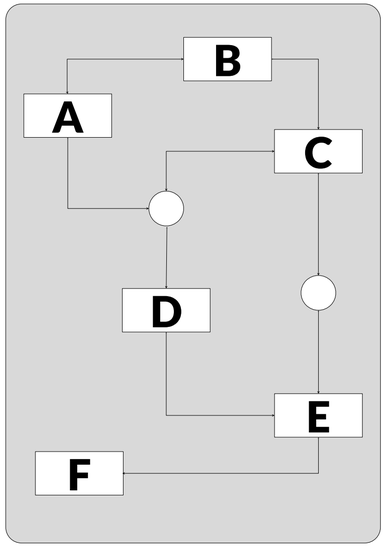
Different types of data

Challenge: Supporting Multiple Tasks

Two central tasks:

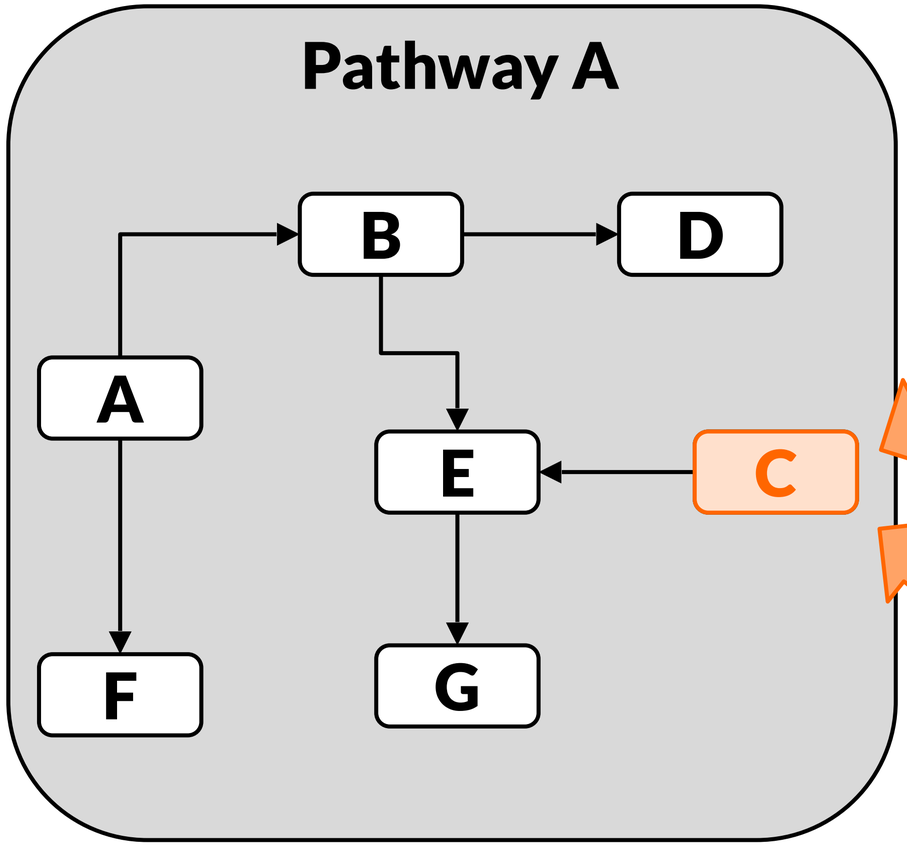
- Explore topology of network
- Explore the attributes of the nodes (experimental data)

Need to support both!



	Sample 1	Sample 2	Sample 3
Gene 1	1	1.1	0.4
Gene 2	2	0.5	1.2
Gene 3	1.4	0.2	0.5
Gene 4	0.3	0.5	0.7

Many Node Attributes

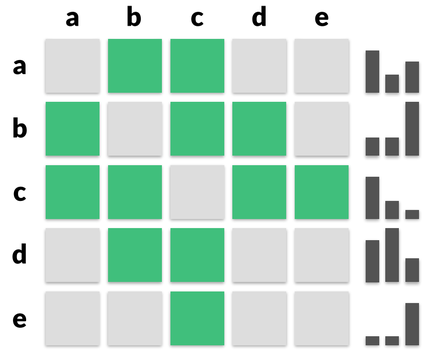


Node	Sample 1	Sample 2	Sample 3	…
A	0.55	0.95	0.83	…
B	0.12	0.42	0.16	…
C	0.33	0.65	0.38	…
…	…	…	…	

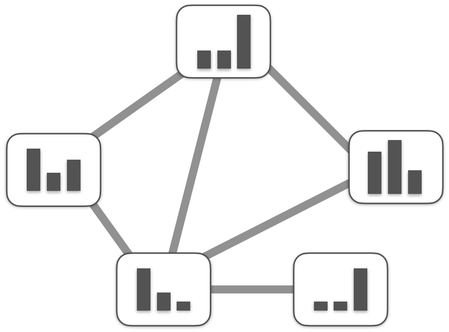
Node	Sample 1	Sample 2	Sample 3	…
A	low	low	very high	…
B	normal	low	high	…
C	high	very low	normal	…
…	…	…	…	

How to visualize attribute data on networks?

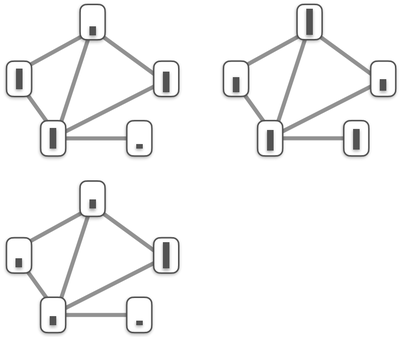
Multivariate Network Visualization Strategies



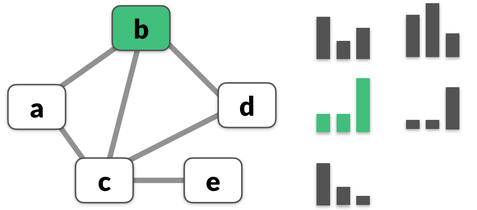
Matrices



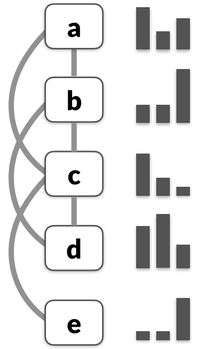
On-Node Encoding



Small Multiples



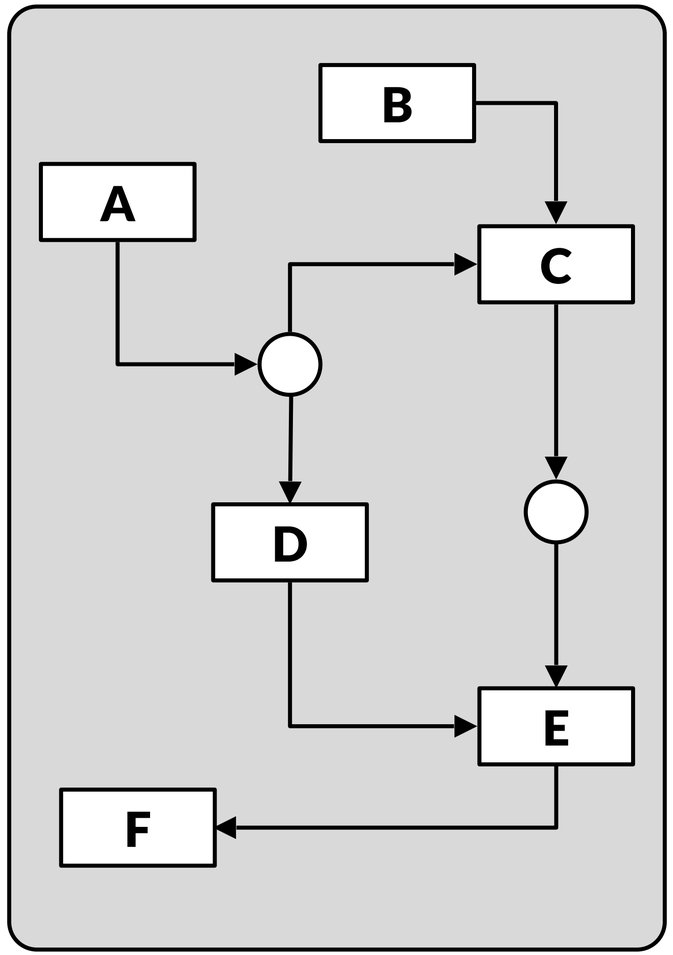
Multiple Coordinated Views



Layout Adaption

Carolina Nobre, Marc Streit, Miriah Meyer, Alexander Lex
The State of the Art in Visualizing Multivariate Networks
Computer Graphics Forum (EuroVis) – https://vdl.sci.utah.edu/mvnv/

Good Old Color Coding



A	-3.4	4.2	5.1	4.2
B	2.8	1.8	1.3	1.1
C	3.1	-2.2	2.4	2.2
D	-3	-2.8	1.6	1.0
E	0.5	0.3	-1.1	1.3
F	0.3	0.3	1.8	-0.3

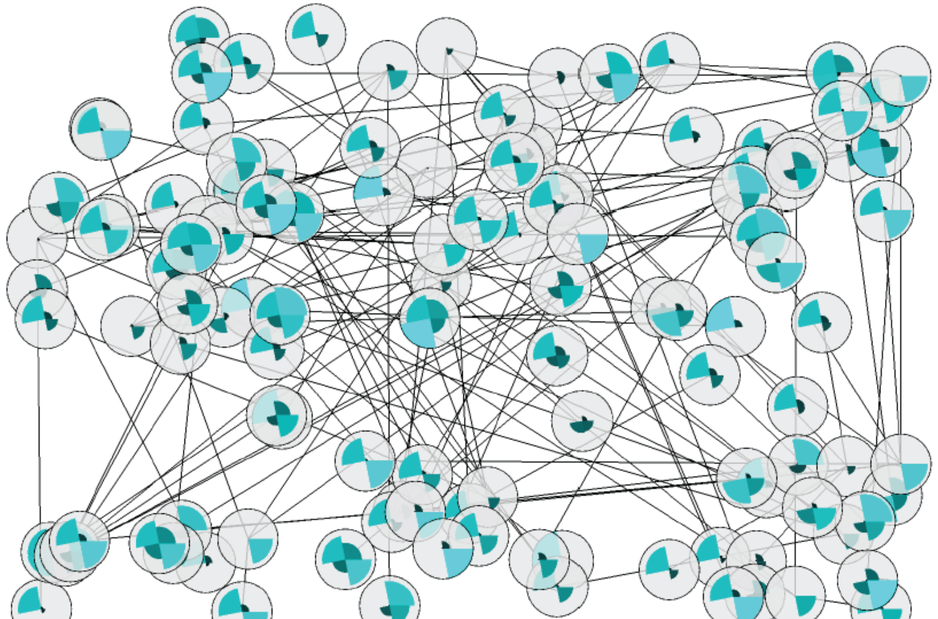
[Lindroos2002]

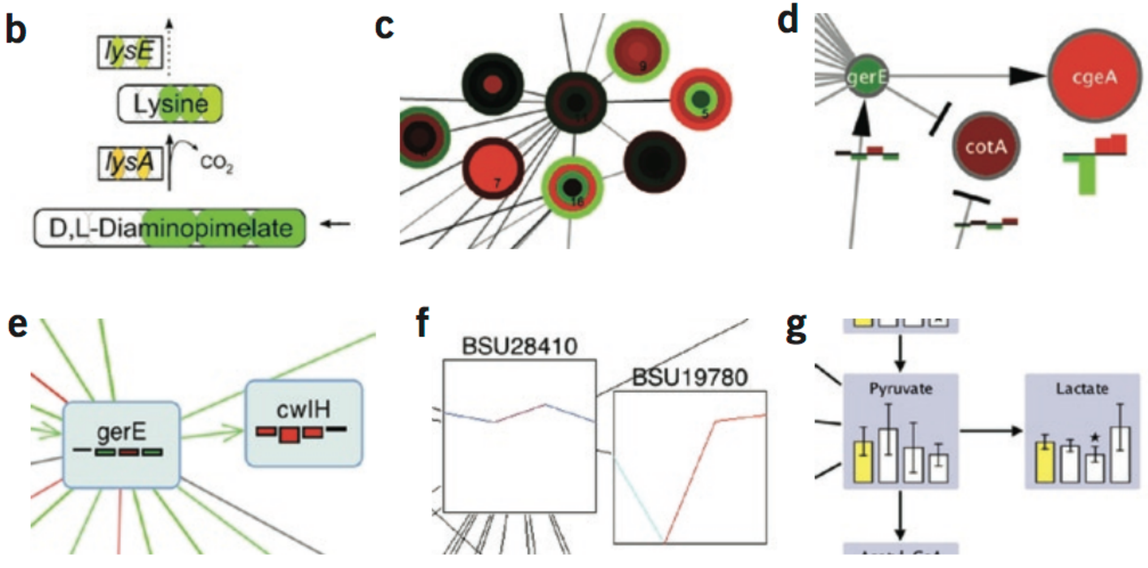
Node Attributes

Coloring

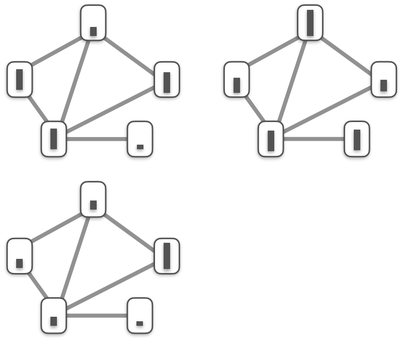
Glyphs

-> Limited in scalability



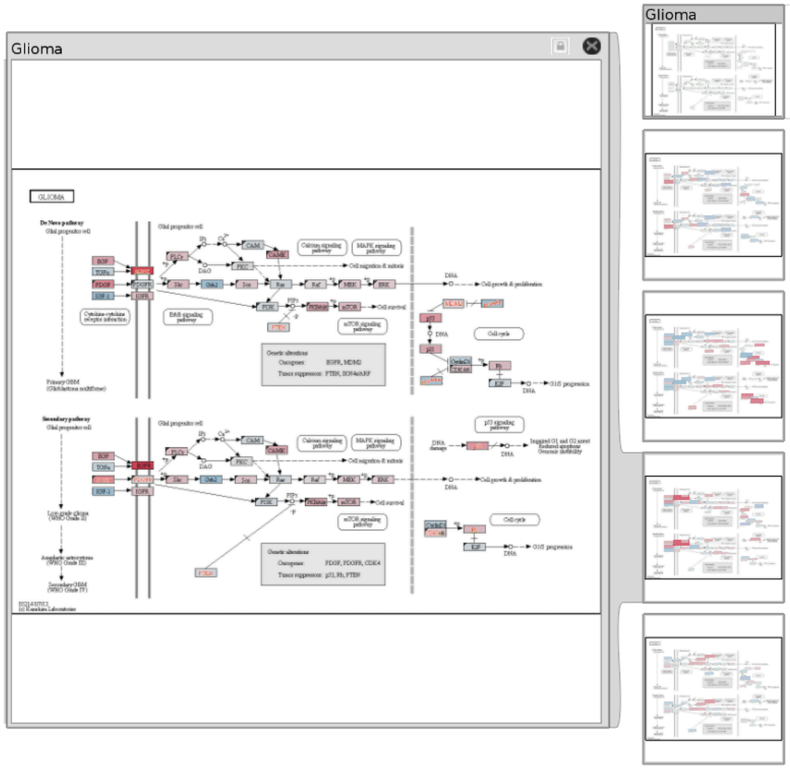


Small Multiples

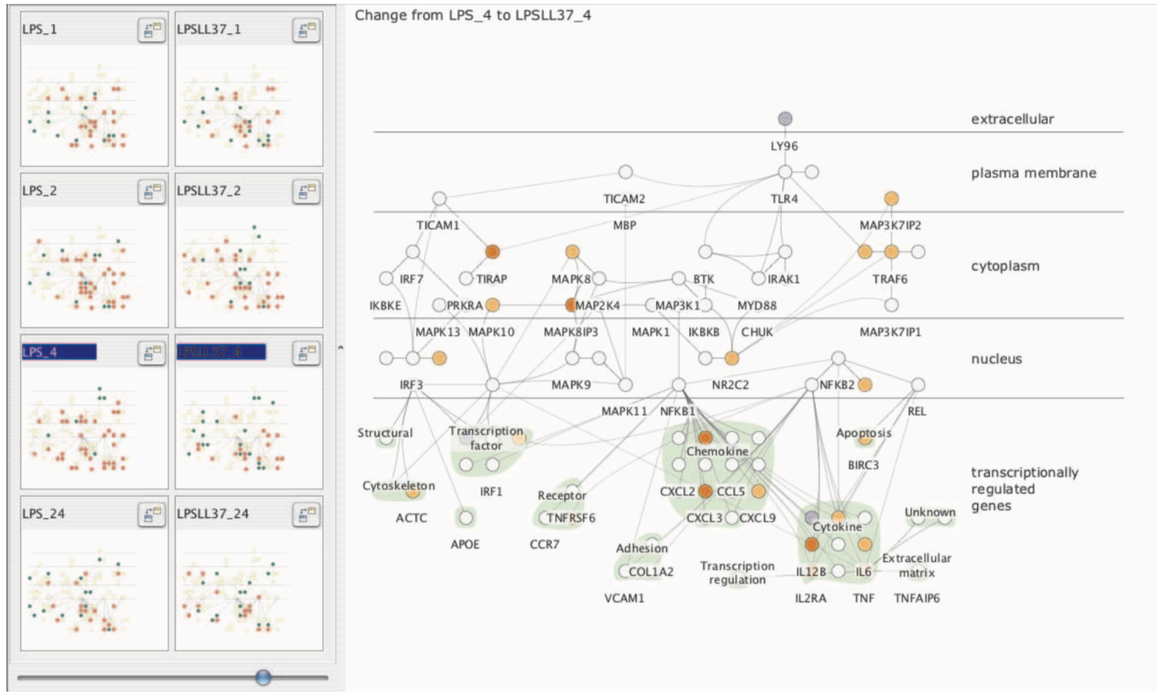


On-node encoding with small multiples

Graphs tend to be small, combine with focus graph

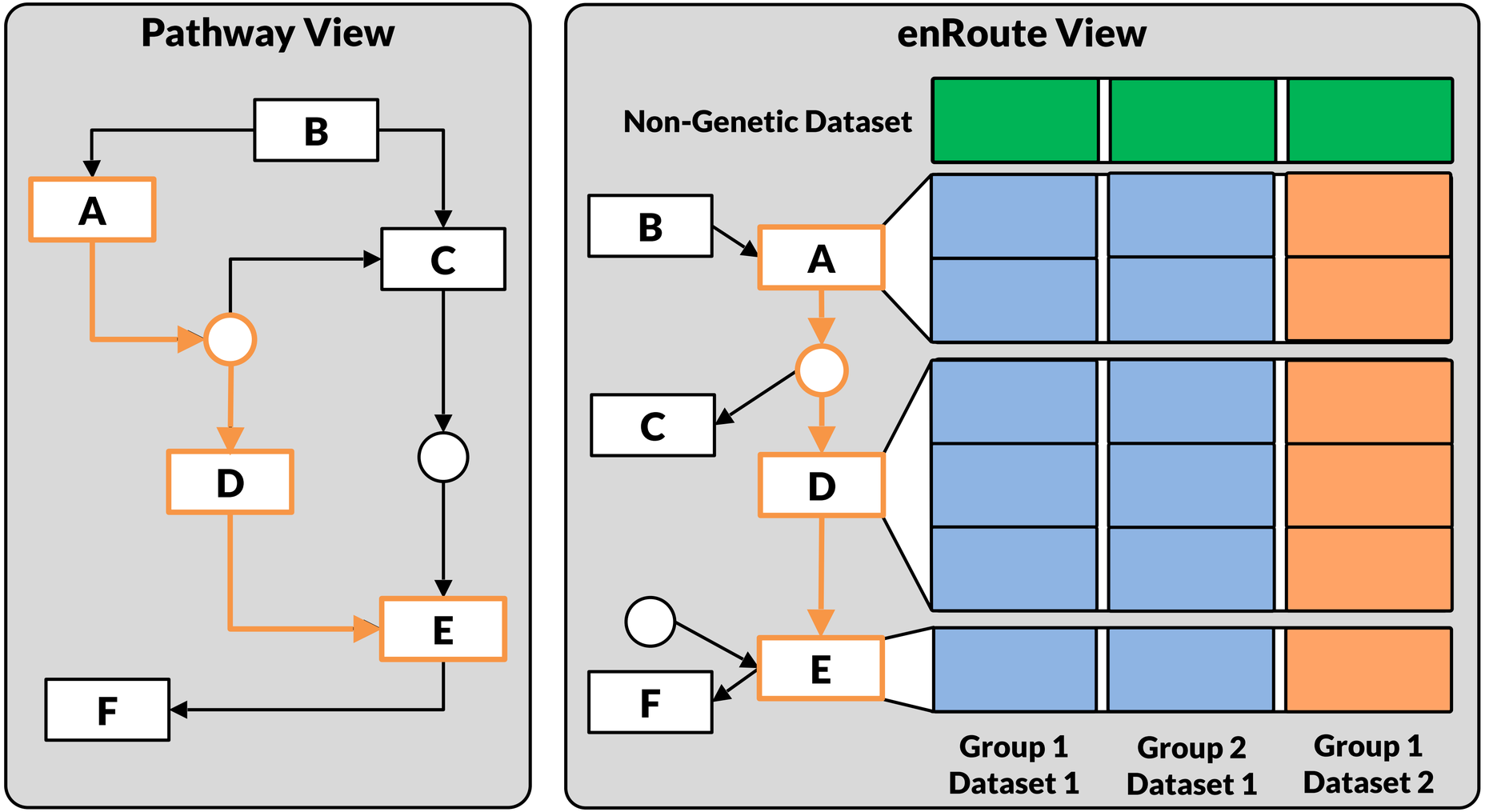


[Lex et al., StratomeX, 2012]

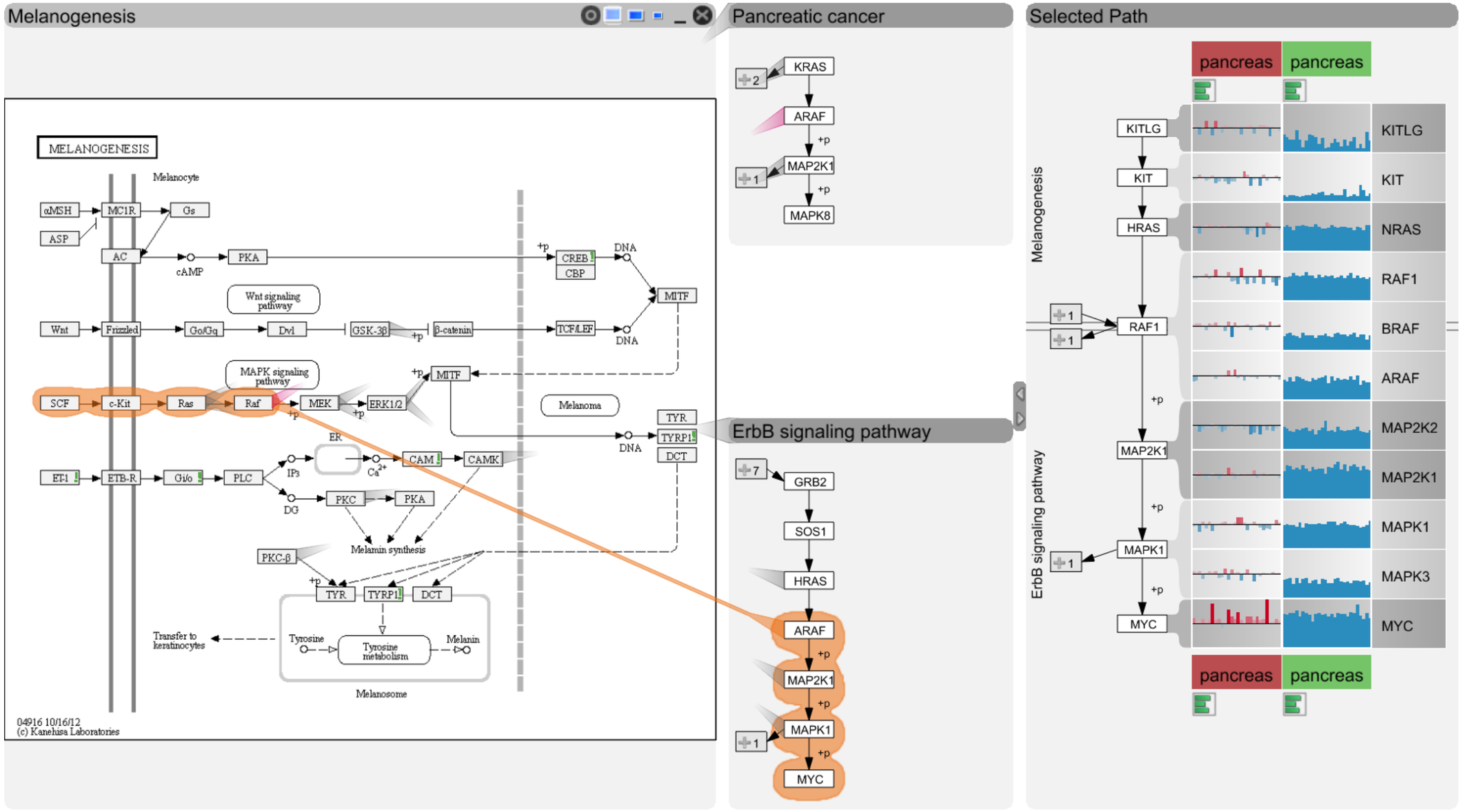


[Barsky et al., Cerebral, 2008]

Path Extraction: enRoute



enRoute


Melanogenesis
MELANOGENESIS
Melanocyte
αMSH
MC1R
Gs
ASP
AC
cAMP
PKA
CREB
CBP
DNA
MITF
Wnt signaling pathway
Wnt
Frizzled
Go/Gq
Dvl
GSK-3β
β-catenin
TCF/LEF
MAPK signaling pathway
SCF
c-Kit
Ras
Raf
MEK
ERK1/2
Melanoma
TYR
TYRP1
DCT
ER
ET1
ETB-R
Gi/o
PLC
IP3
Ca2+
CAM
CAMK
DG
PKC
PKA
Melamin synthesis
PKC-β
Transfer to keratinocytes
Tyrosine
Tyrosine metabolism
Melanin
Melanosome
04916 10/16/12
(c) Kanehisa Laboratories
Pancreatic cancer
KRAS
ARAF
MAP2K1
MAPK8
ErbB signaling pathway
GRB2
SOS1
HRAS
ARAF
MAP2K1
MAPK1
MYC
Selected Path
pancreas
pancreas
Melanogenesis
ErbB signaling pathway
KITLG
KIT
HRAS
RAF1
MAP2K1
MAPK1
MYC
NRAS
BRAF
ARAF
MAP2K2
MAP2K1
MAPK1
MAPK3
MYC
pancreas
pancreas

File Data Window View Help

Entourage

Pathways

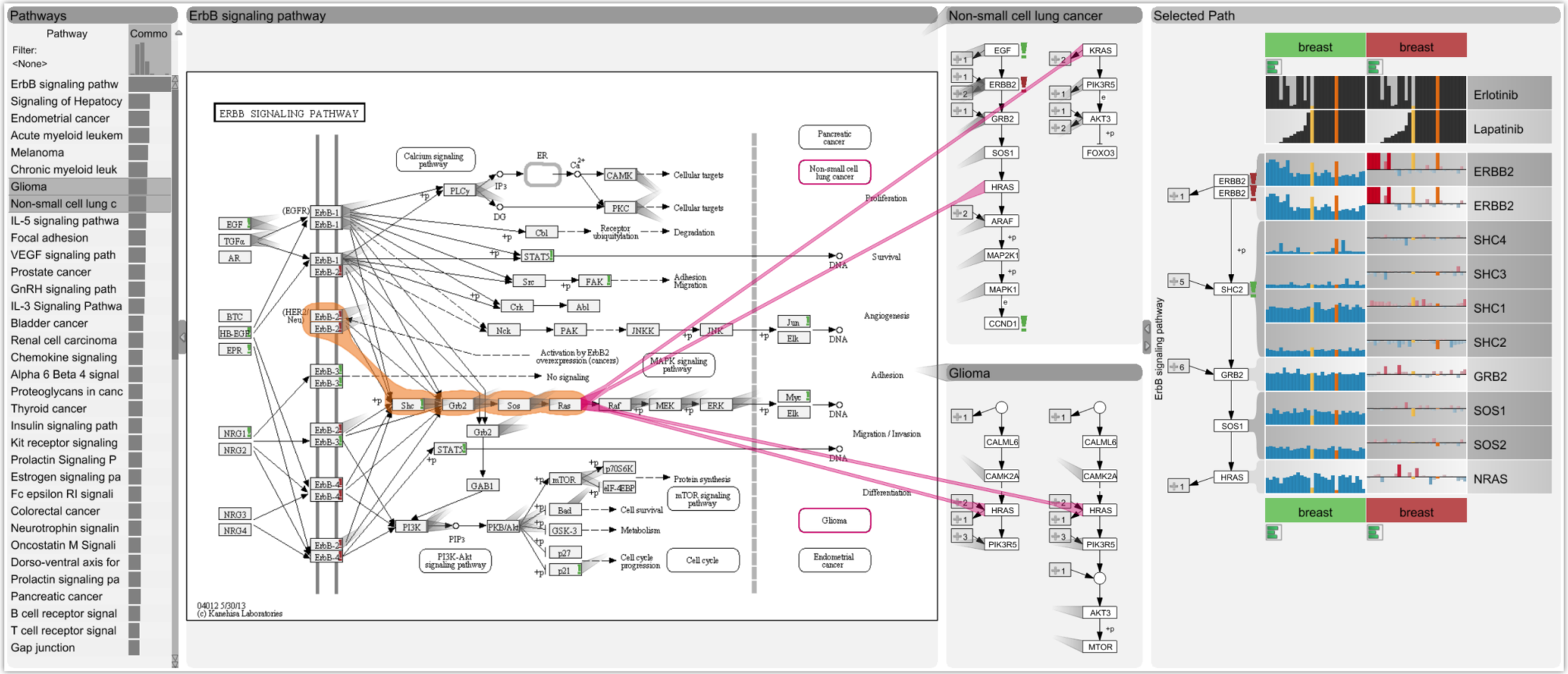
Pathway

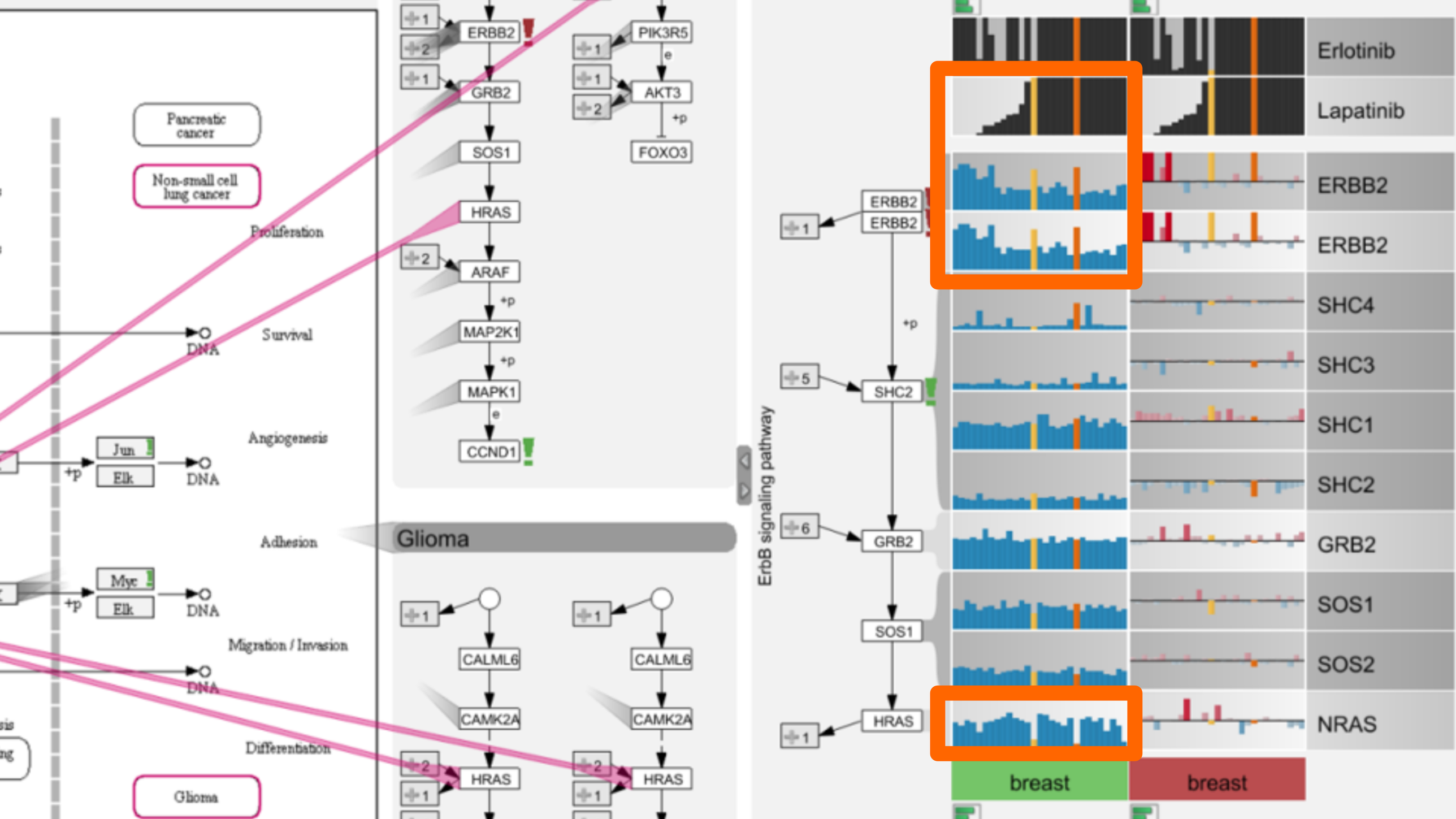
Filter:
<None>

1 C donor
2-Oxocarboxylic acid
ABC transporters
ABC-family proteins
ACE Inhibitor Pathwa
Acetylcholine Synthes
Acute myeloid leukem
Adherens junction
Adipocyte TarBase
Adipocytokine signali
Adipogenesis
Advanced glycosylatio
Aflatoxin B1 metaboli
African trypanosomias
AGE/RAGE pathway
AhR pathway
Alanine and aspartate
Alanine, aspartate an
Alcoholism
Aldosterone-regulated
Allograft rejection
Allograft rejection
Alpha 6 Beta 4 signal
alpha-Linolenic acid
Alzheimer's disease
Alzheimers Disease
amino acid conjugatio
amino acid conjugatio
Amino sugar and nucl
Aminoacyl-tRNA bios
Amoebiasis
Amphetamine addicti
AMPK signaling
Amyotrophic lateral sc
Androgen receptor si
Angiogenesis
Angiogenesis
angiogenesis overvie
Antigen processing an
APC/C-mediated degra
Apoptosis
Apoptosis
Apoptosis Meta Path
Apoptosis Modulation
Apoptosis Modulation
Apoptosis, anoikis an

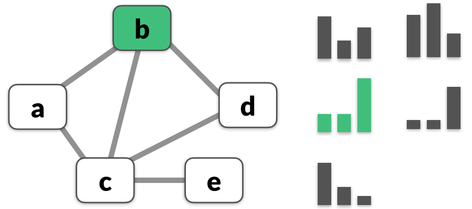
Selected Path

Case Study: CCLE Data



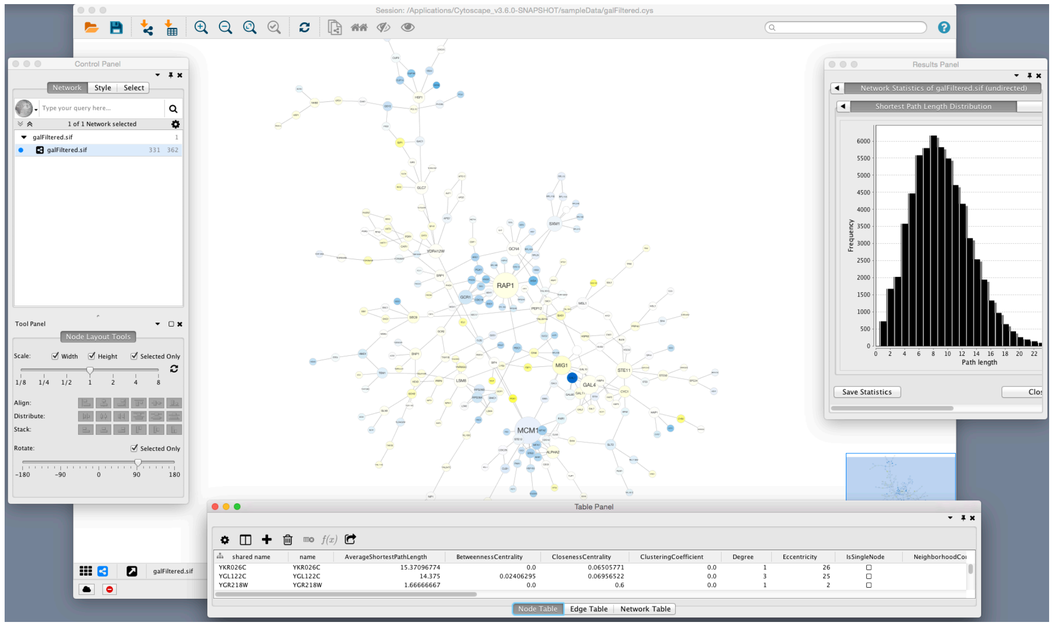

Pancreatic cancer
Non-small cell lung cancer
Proliferation
Survival
Angiogenesis
Adhesion
Migration / Invasion
Differentiation
Glioma
DNA
Jun
Elk
Myc
+p
ERBB2
GRB2
SOS1
HRAS
ARAF
MAP2K1
MAPK1
CCND1
PIK3R5
AKT3
FOXO3
CALML6
CAMK2A
ErbB signaling pathway
SHC2
Erlotinib
Lapatinib
ERBB2
SHC4
SHC3
SHC1
SHC2
GRB2
SOS1
SOS2
NRAS
breast
breast

Multiple Coordinated Views

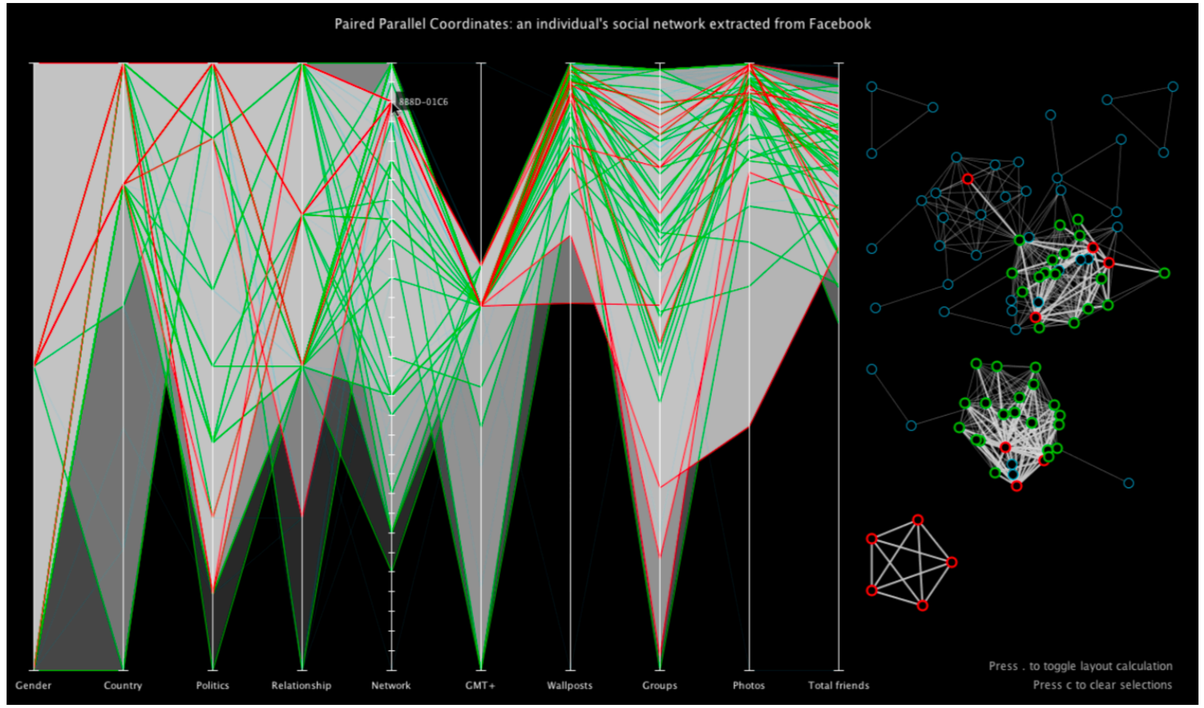


Can optimize for topology and attributes at the same time

Lacking whith regards to interplay



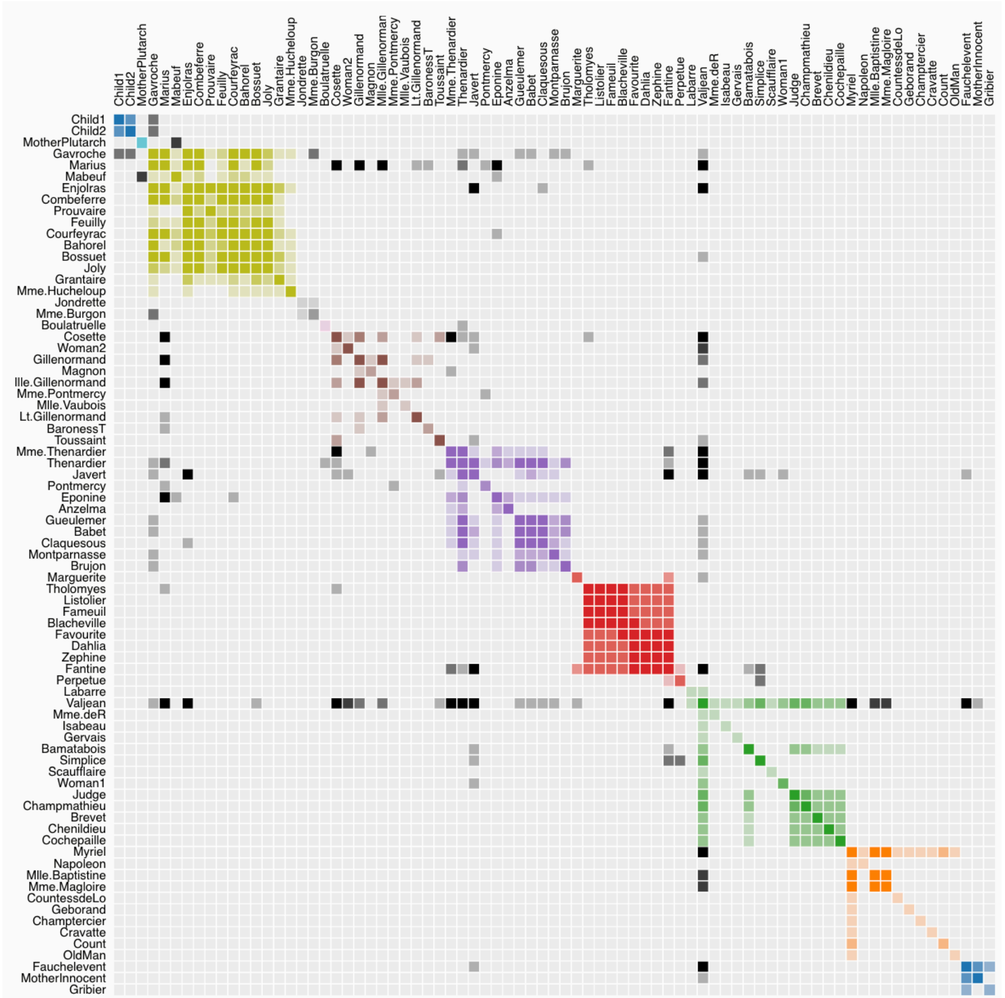
[Shannon et al., Cytoscape, 2003]



[Shannon et al., 2008]

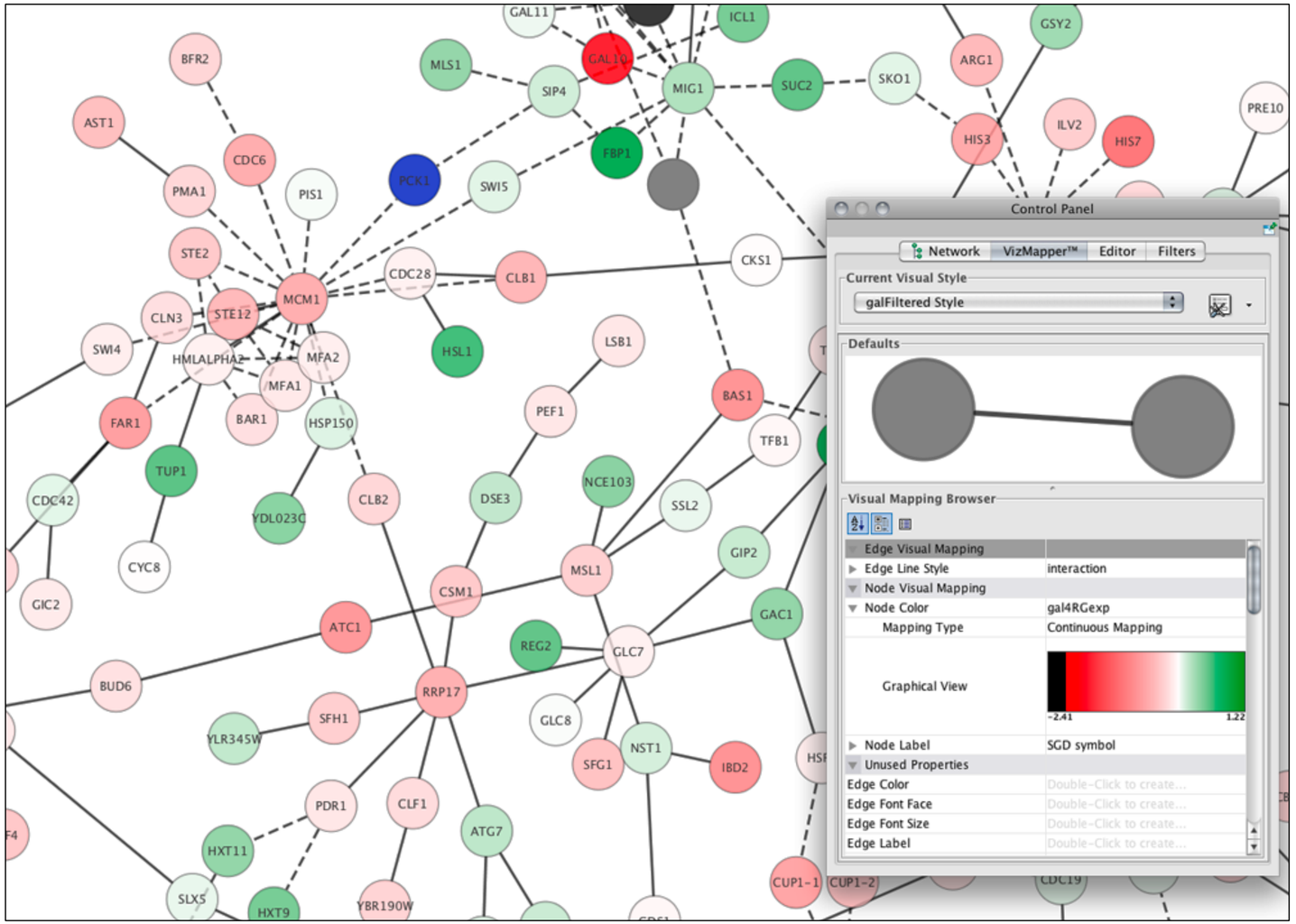
Visualizing Edge Attributes

Good Choice: Matrix



Cytoscape

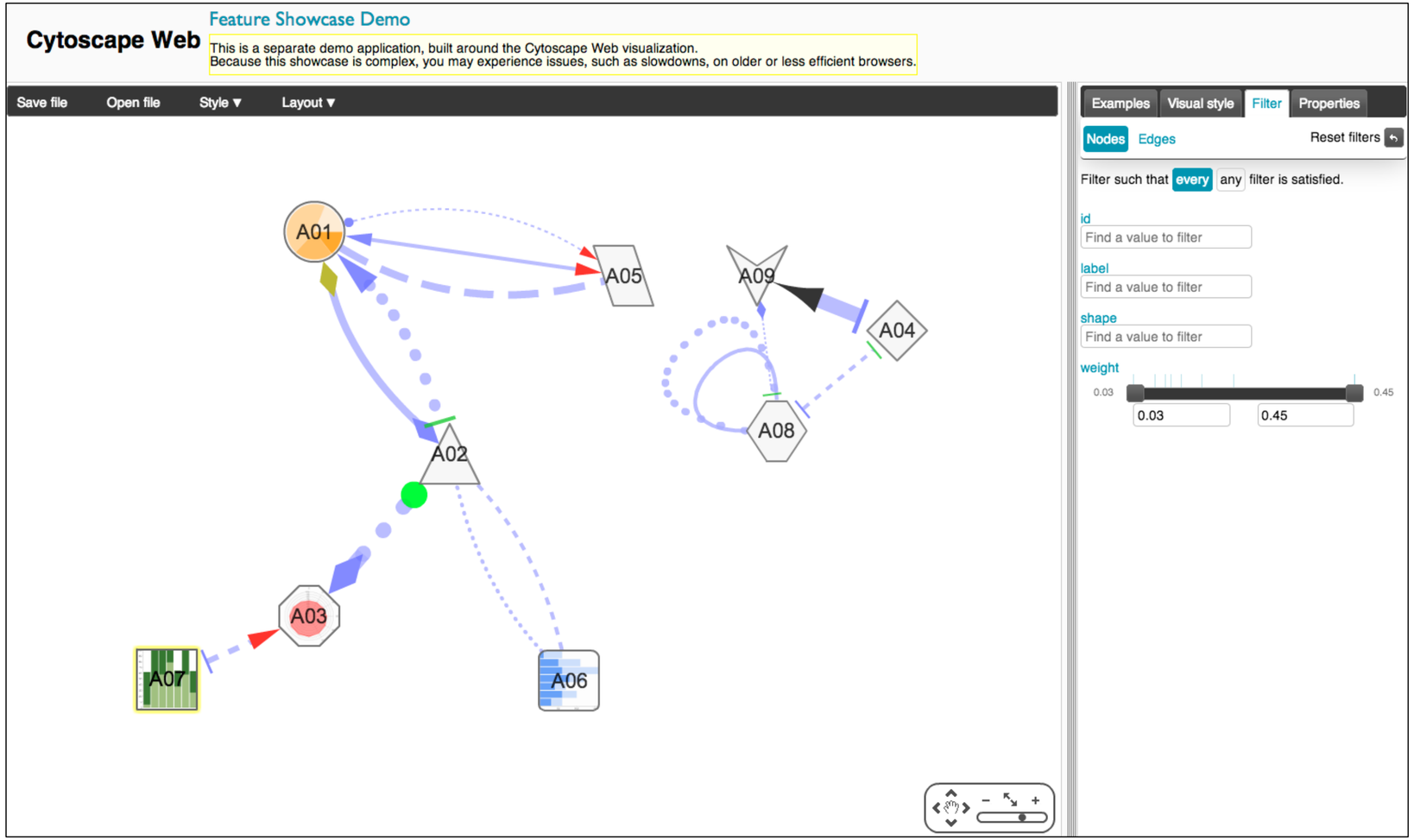
Open source platform for complex network analysis



http://www.cytoscape.org/

Cytoscape Web

http://cytoscapeweb.cytoscape.org/



Alexander Lex
@alexander_lex
http://alexander-lex.net

THE UNIVERSITY OF UTAH

Questions?

visualization design lab

SCI
www.sci.utah.edu