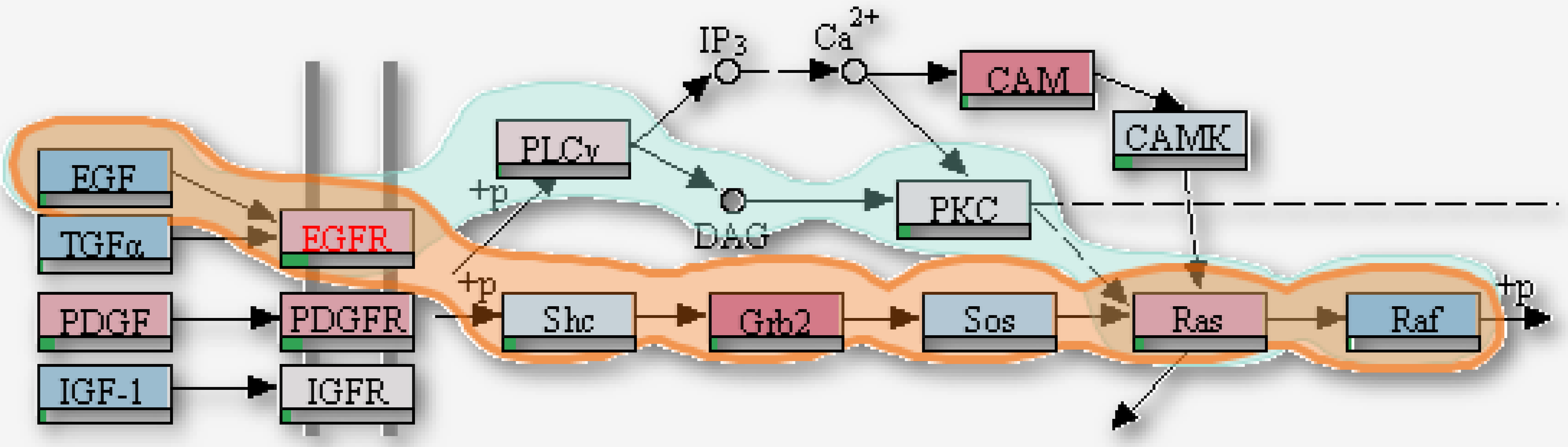
Visualization Approaches for Biomolecular Data

Alexander Lex



School of Interactive Computing
Georgia Tech

April 8, 2014

CALEYDO

HARVARD
School of Engineering
and Applied Sciences

Who am I?

alexander-lex.com

PostDoc @ Harvard,
Hanspeter Pfister's Group

PhD from TU Graz, Austria

Co-leader of
Caleydo Project

What is CALEYDO ?

Software for analyzing molecular biology data

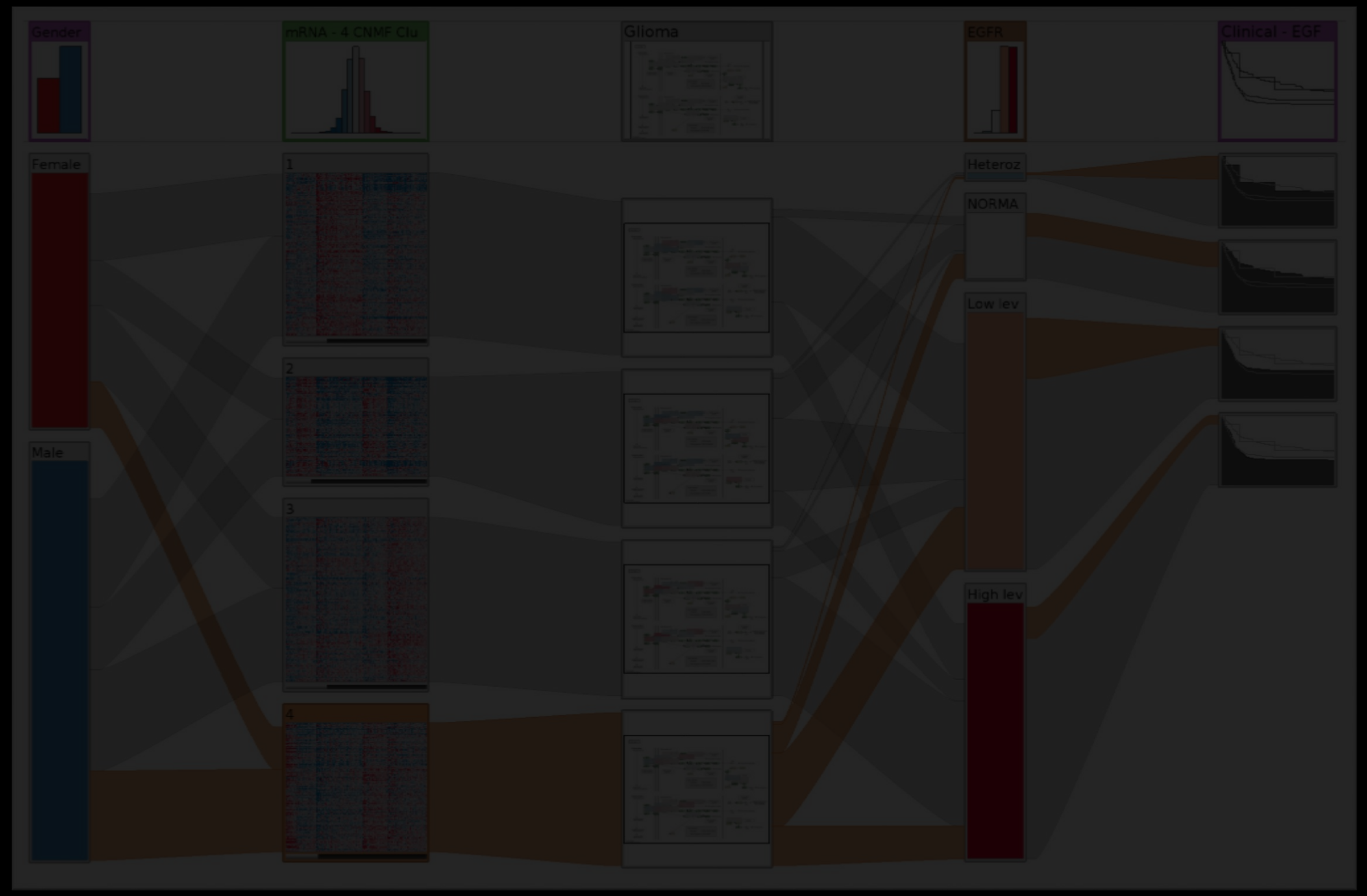
Software for doing research in visualization

Quest for compromise between
academic prototyping and ready-to-use software

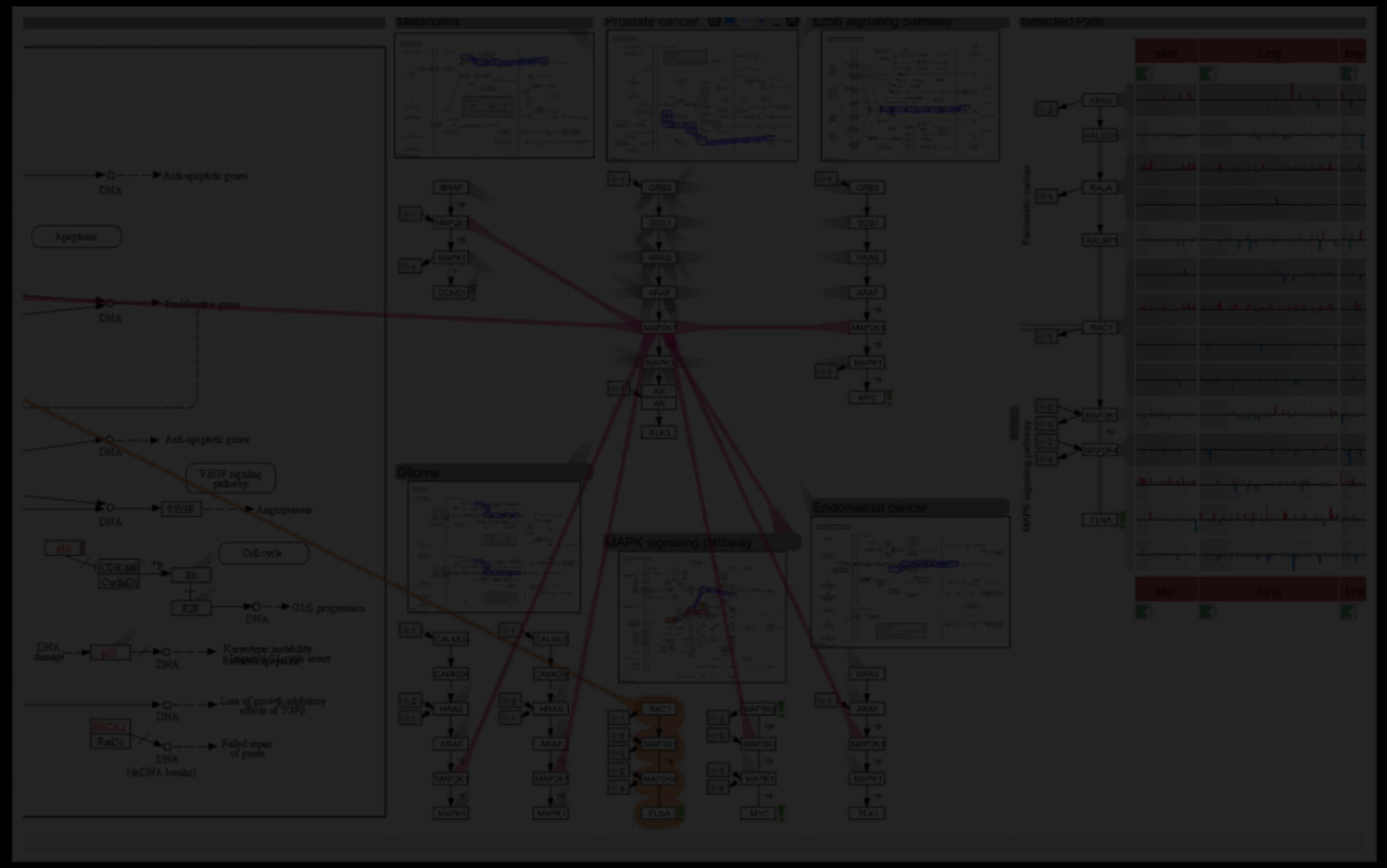
http://caleydo.org

The Team

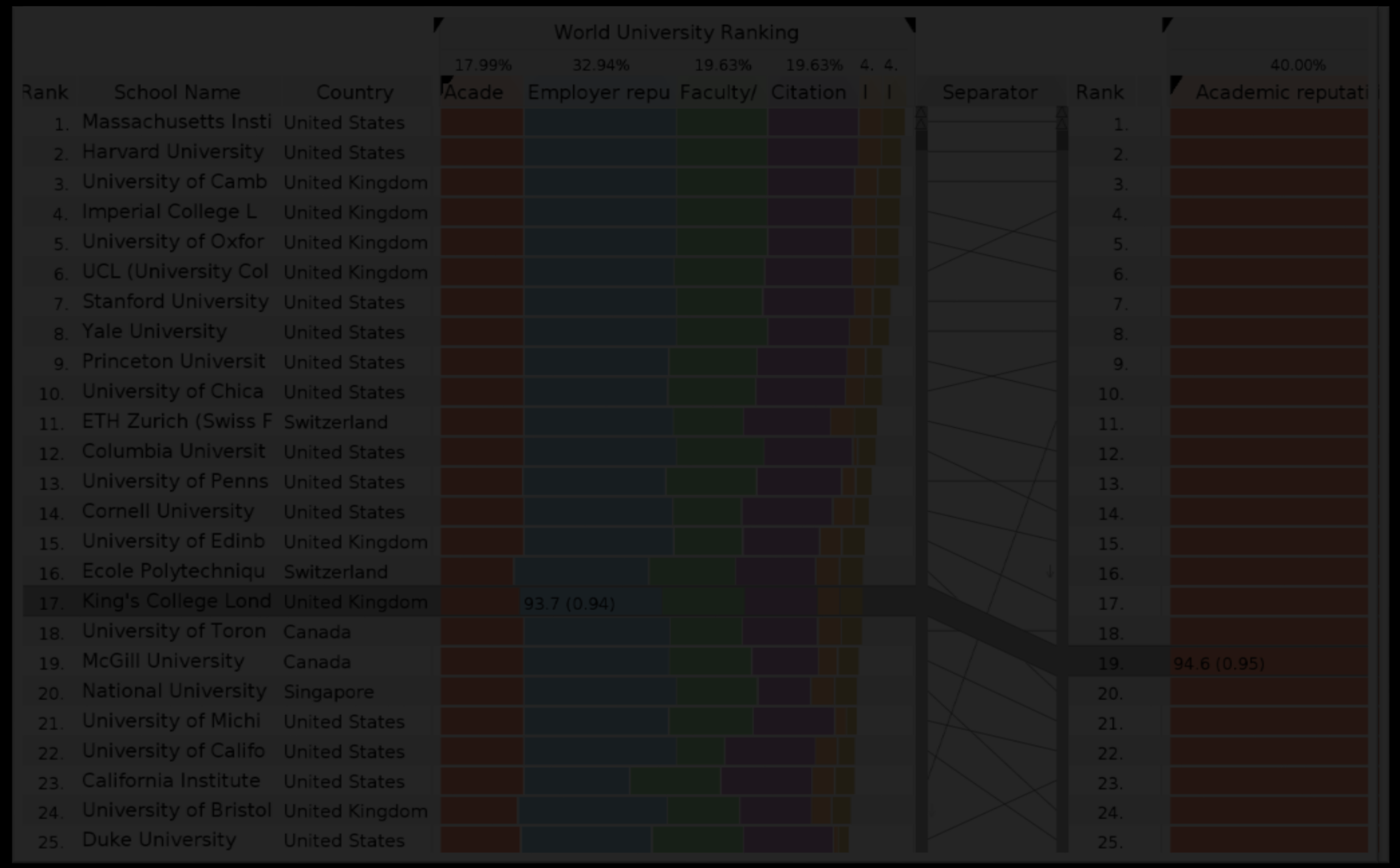
Marc Streit	*Johannes Kepler University Linz, AT*
Christian Partl	*Graz University of Technology, AT*
Samuel Gratzl	*Johannes Kepler University Linz, AT*
Nils Gehlenborg	*Harvard Medical School, Boston, USA*
Dieter Schmalstieg	*Graz University of Technology, AT*
Hanspeter Pfister	*Harvard University, Cambridge, USA*



Heterogeneous Datasets

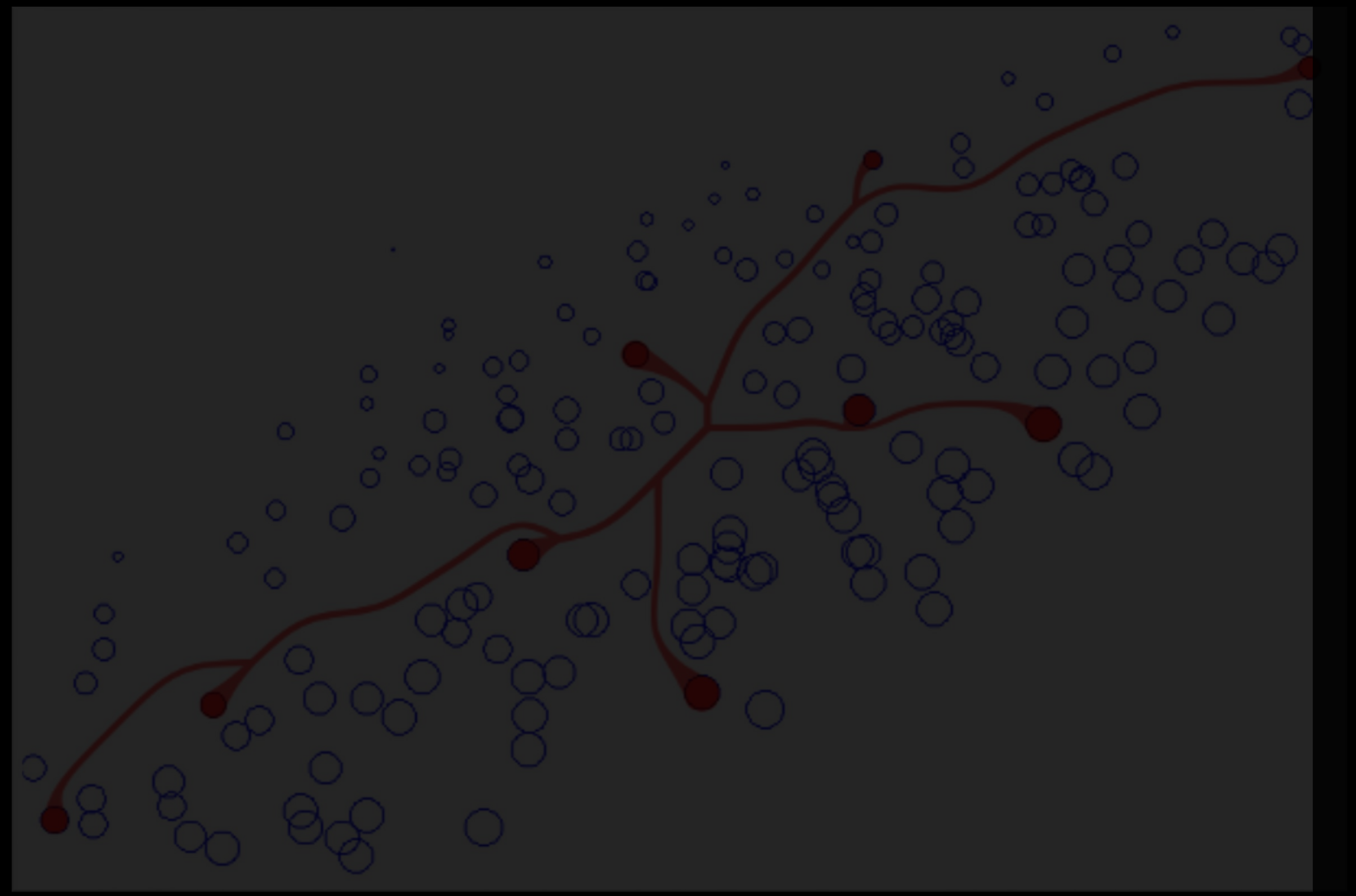


Multi-Variate Graphs



Multi-Attribute Rankings

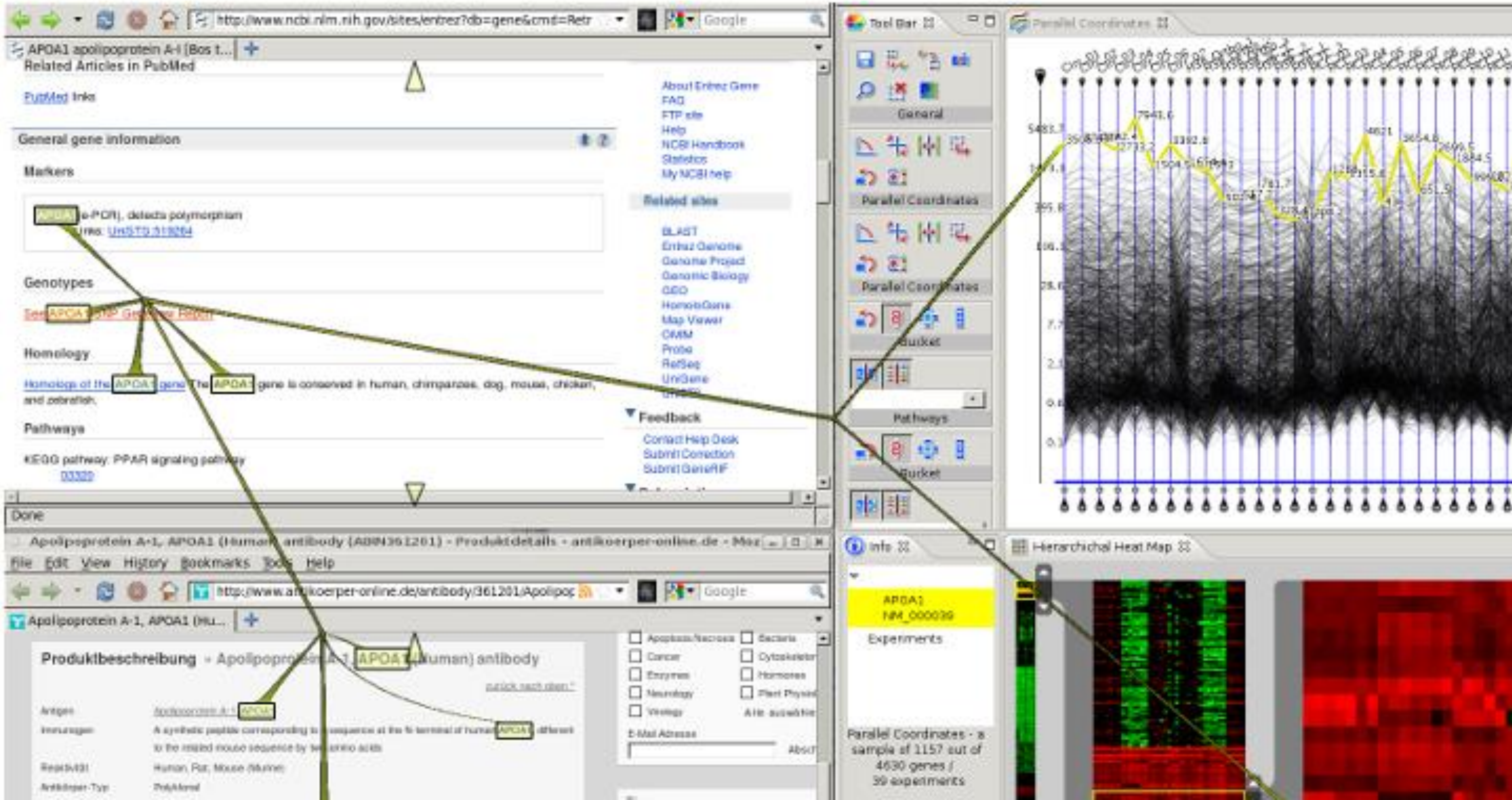
Multi-Dim. Datasets



Visual Linking

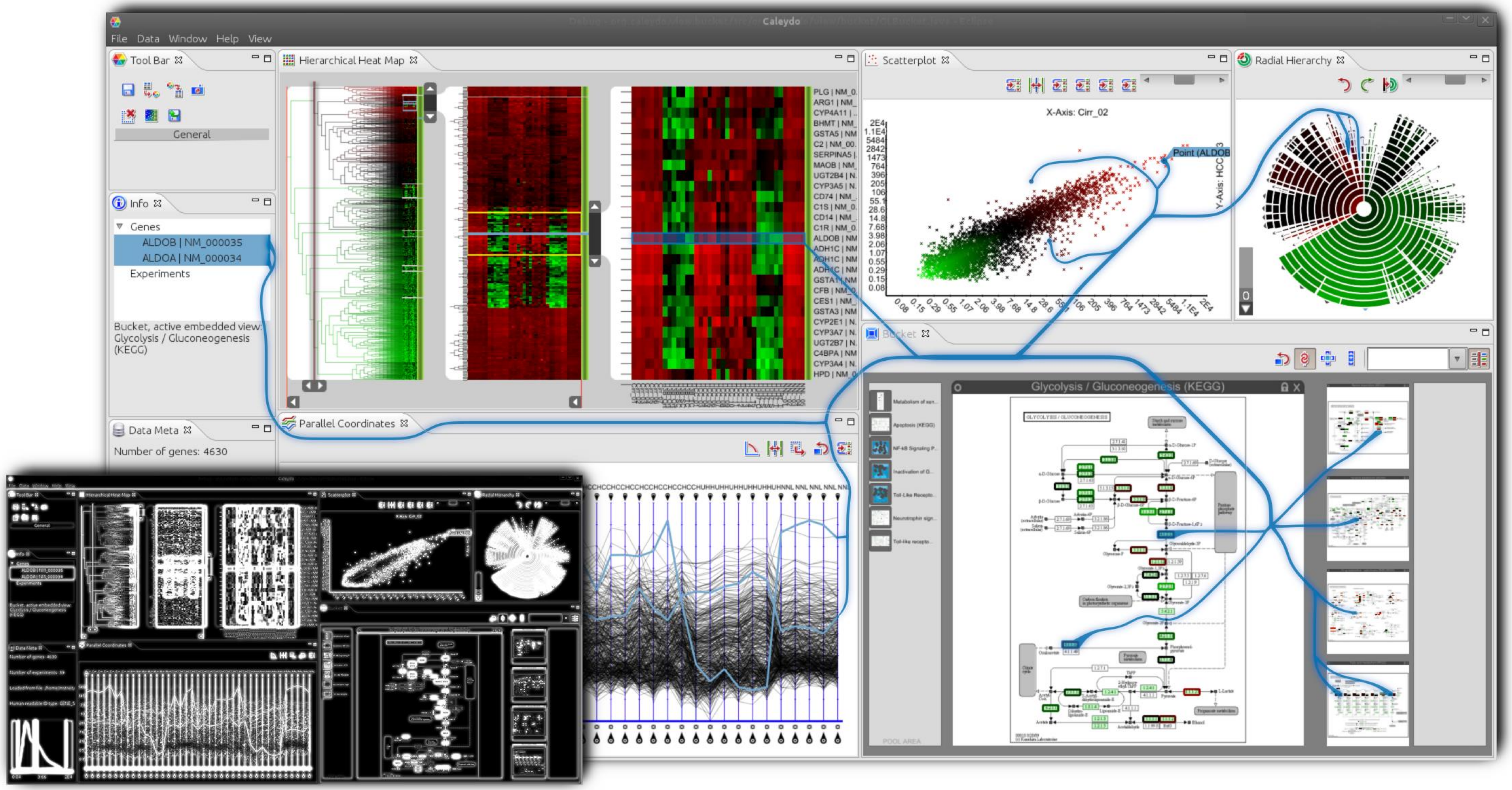
Set Visualization

Visual Linking Across Applications



[Waldner GI'10]
Best Paper Award

Context Preserving Visual Links



[Steinberger InfoVis’11]
Best Paper Award

Visualizing Hidden Content

[Geymayer CHI'14]
Honorable Mention Award

[Lex, EuroVis '12]
3rd Best Paper Award

Heterogeneous Data
Cancer Subtypes

biology research requires understanding data

but there is so much of it...

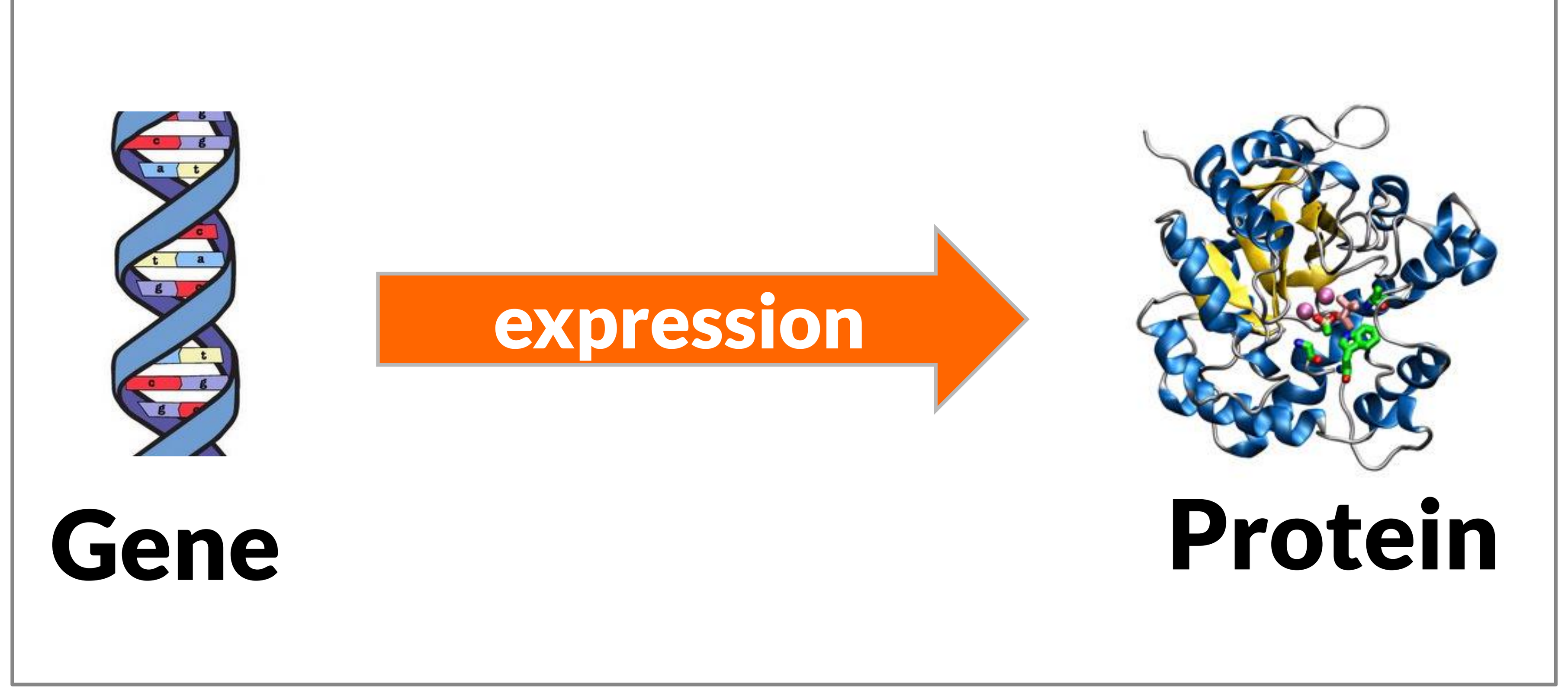
Data Visualization

... makes data accessible

... combines strengths of humans and computers

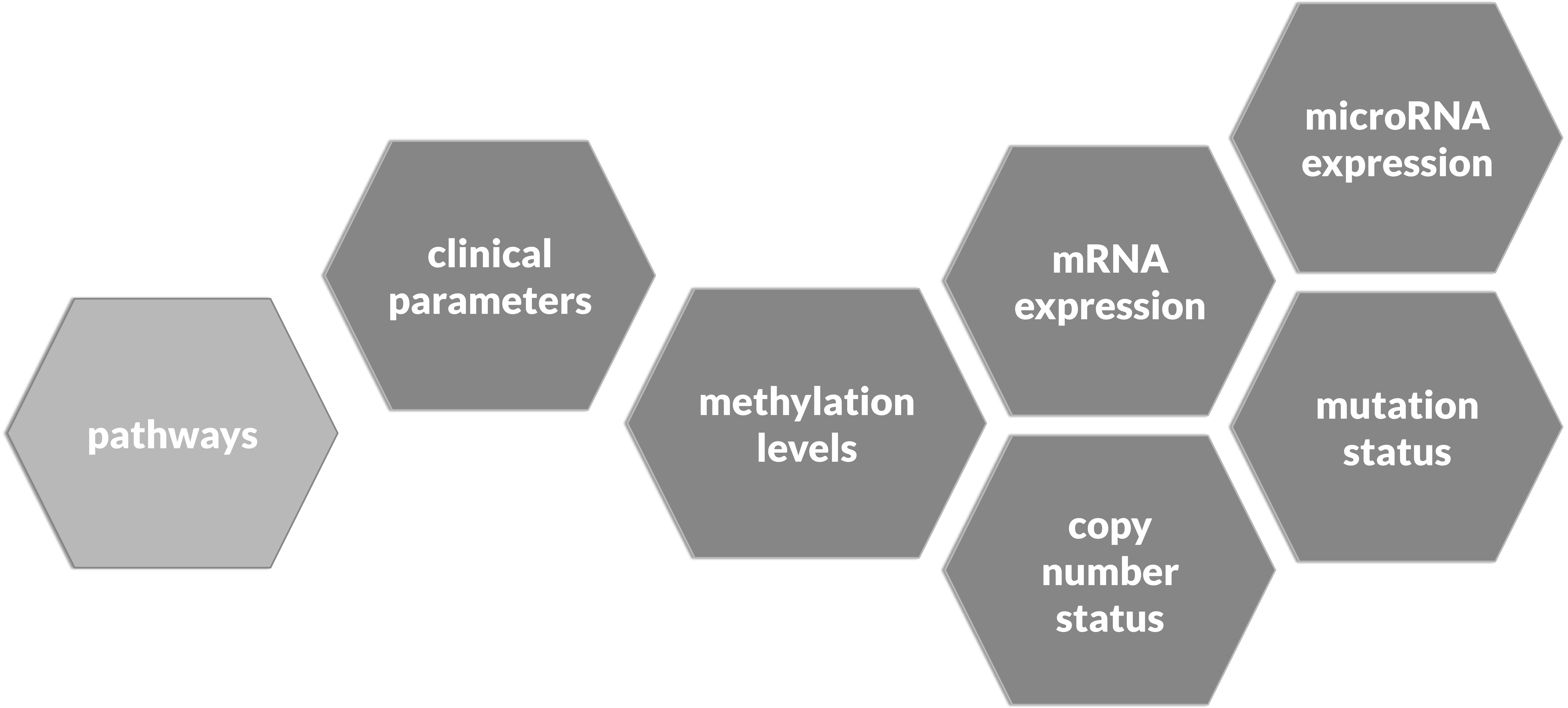
... enables insight

... communicates



The Cancer Genome Atlas

Understanding genomics to improve cancer care



Cancer Subtypes

Cancer is not homogeneous

- different histology
- different molecular alterations

Subtypes have serious implications

- different treatment for subtypes
- prognosis varies between subtypes

Cancer Subtype Analysis

many different types of data

for

large numbers of patients.

StratomeX visualizes…

… the relationships between multiple heterogeneous datasets

… the data within the datasets

… alternative clusterings & groupings

… the effect of groupings on clinical parameters & biological processes

Stratifying Patients

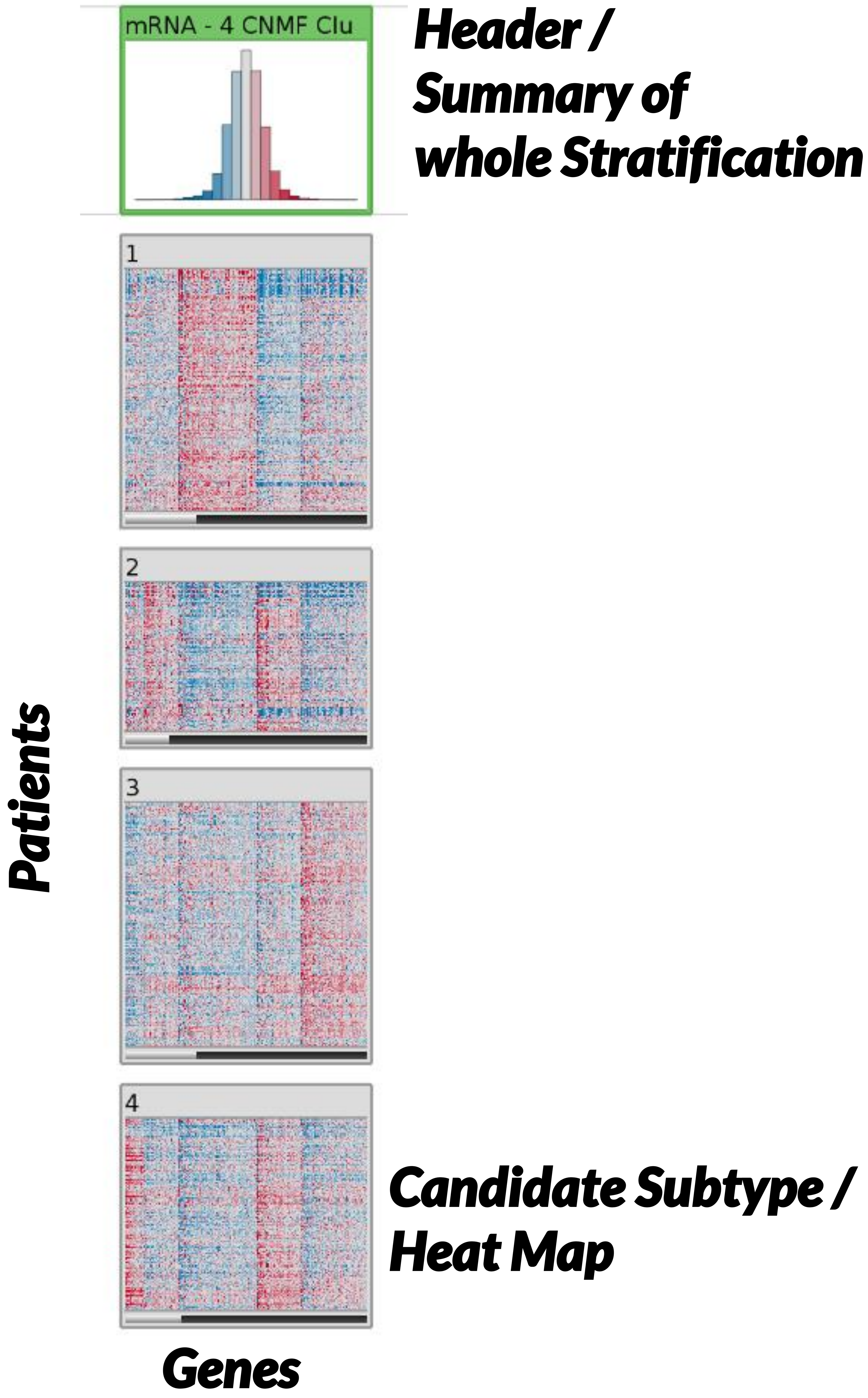
Cluster A1

Cluster A2

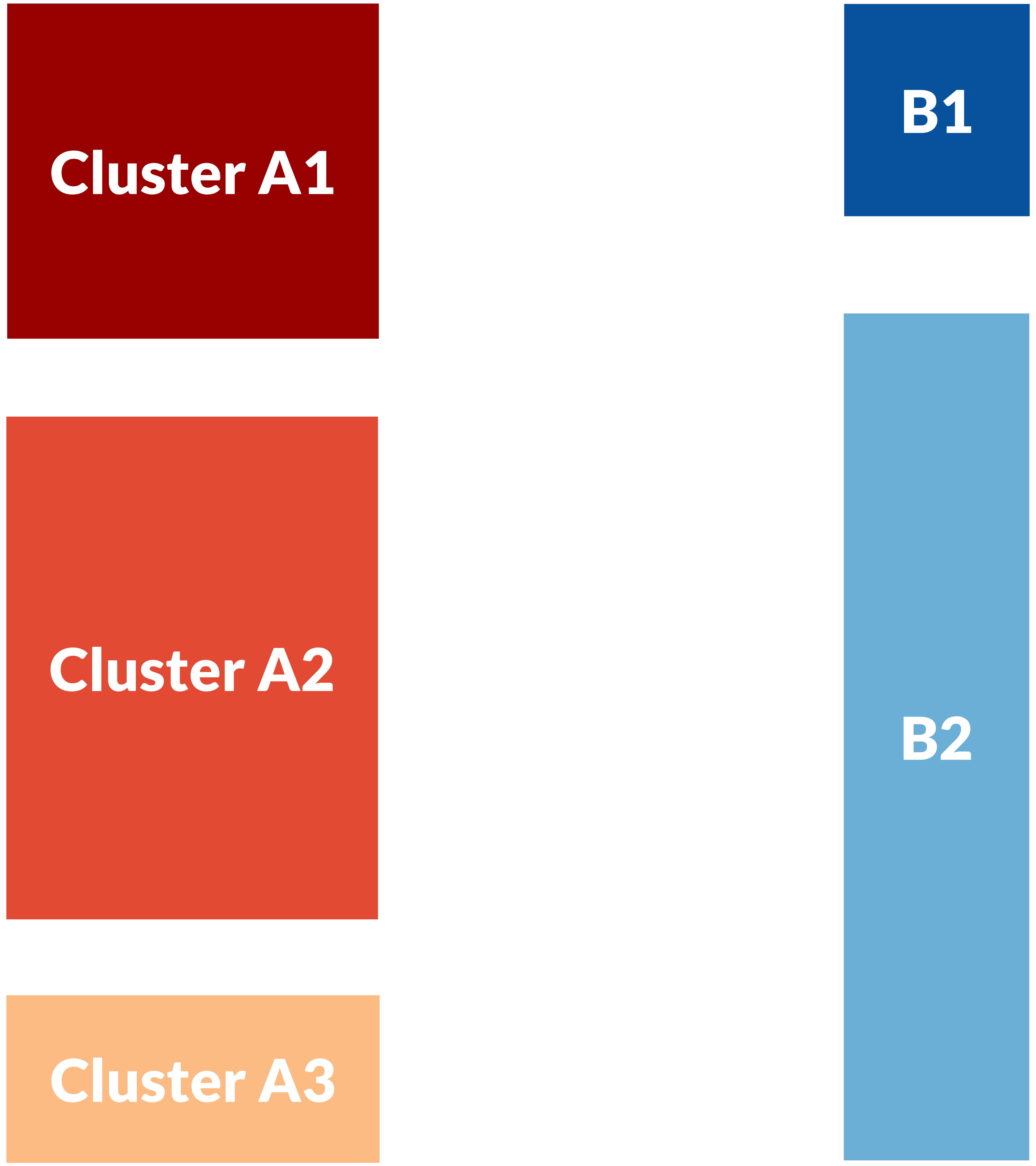
Cluster A3

Subtypes are identified by stratifying datasets, e.g.,

- based on an expression pattern
- a mutation status
- a copy number alteration
- a combination of these


mRNA - 4 CNMF Clu
Header /
Summary of
whole Stratification
1
2
3
4
Patients
Candidate Subtype /
Heat Map
Genes

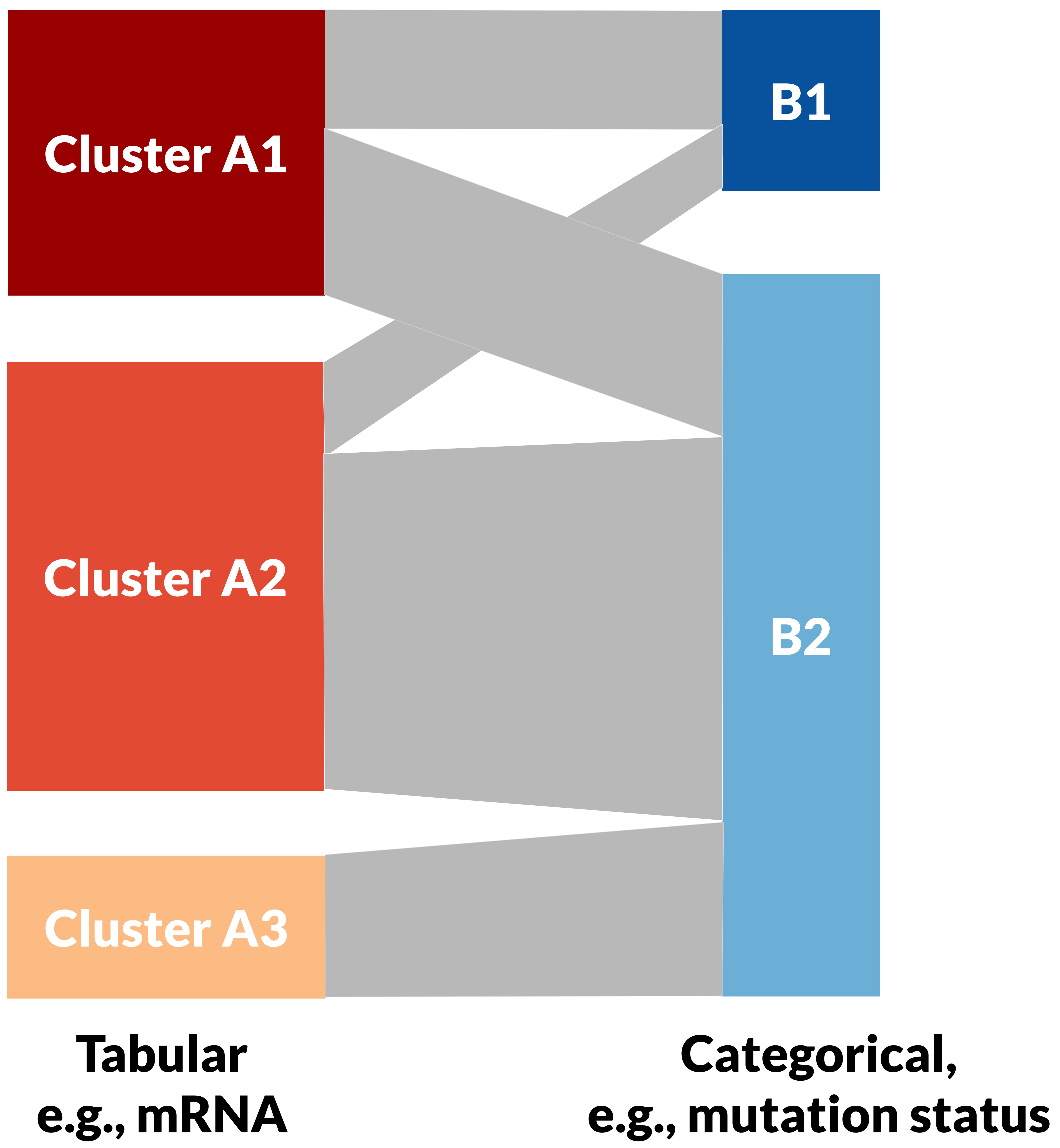
Comparing Stratifications



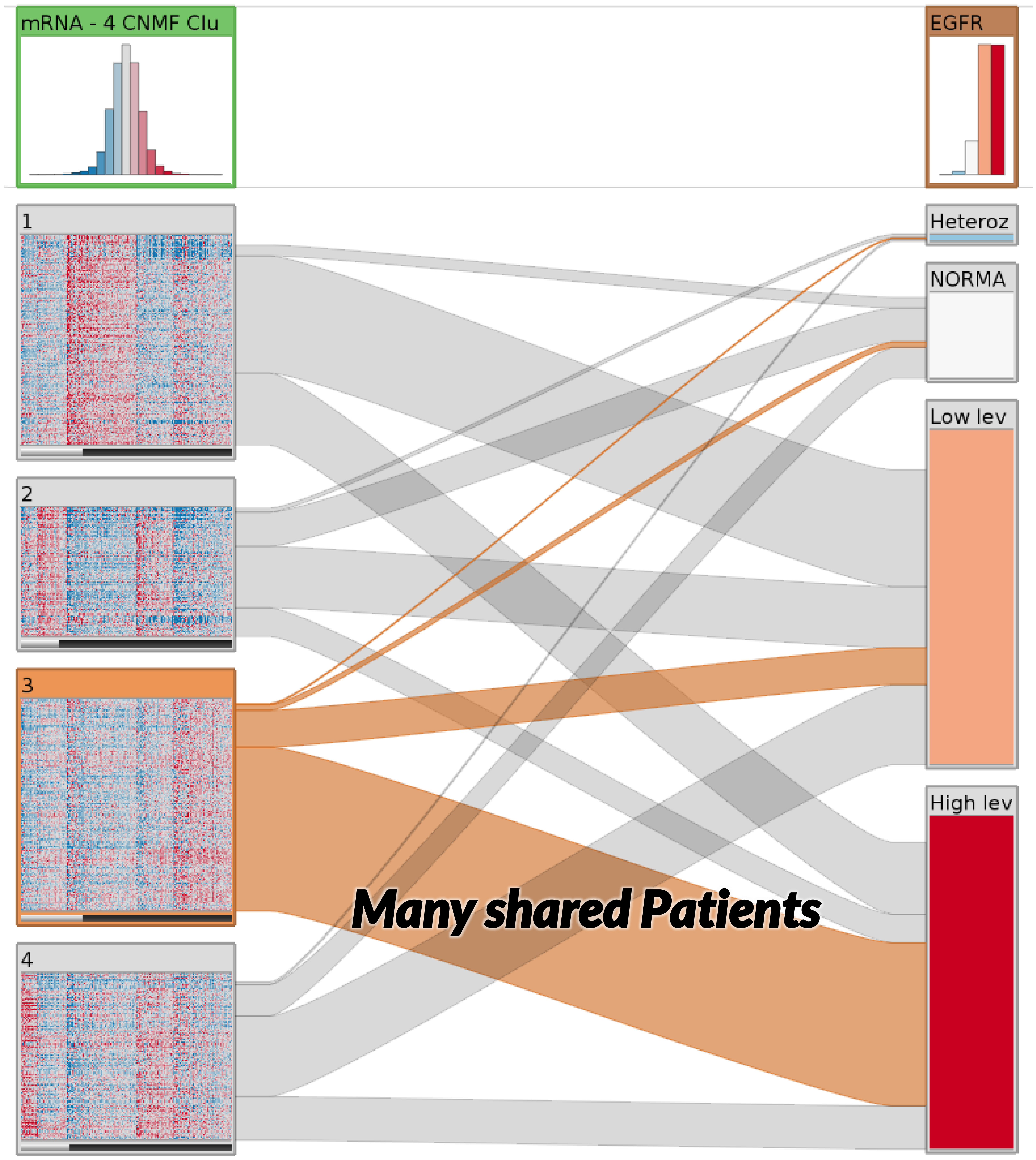
Tabular
e.g., mRNA

Categorical,
e.g., mutation status

Comparing Stratifications



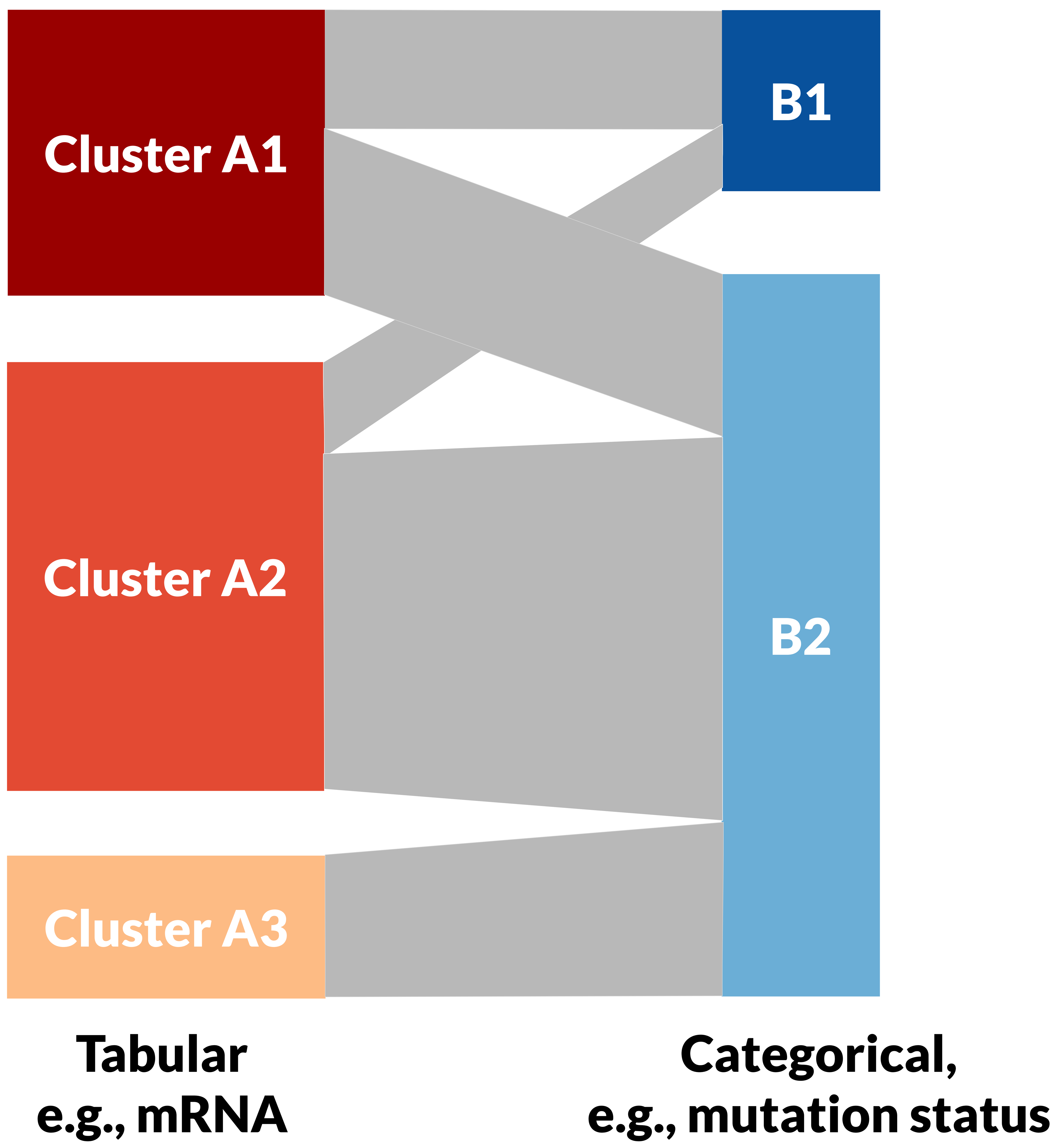
[Parallel Sets, Kosara]



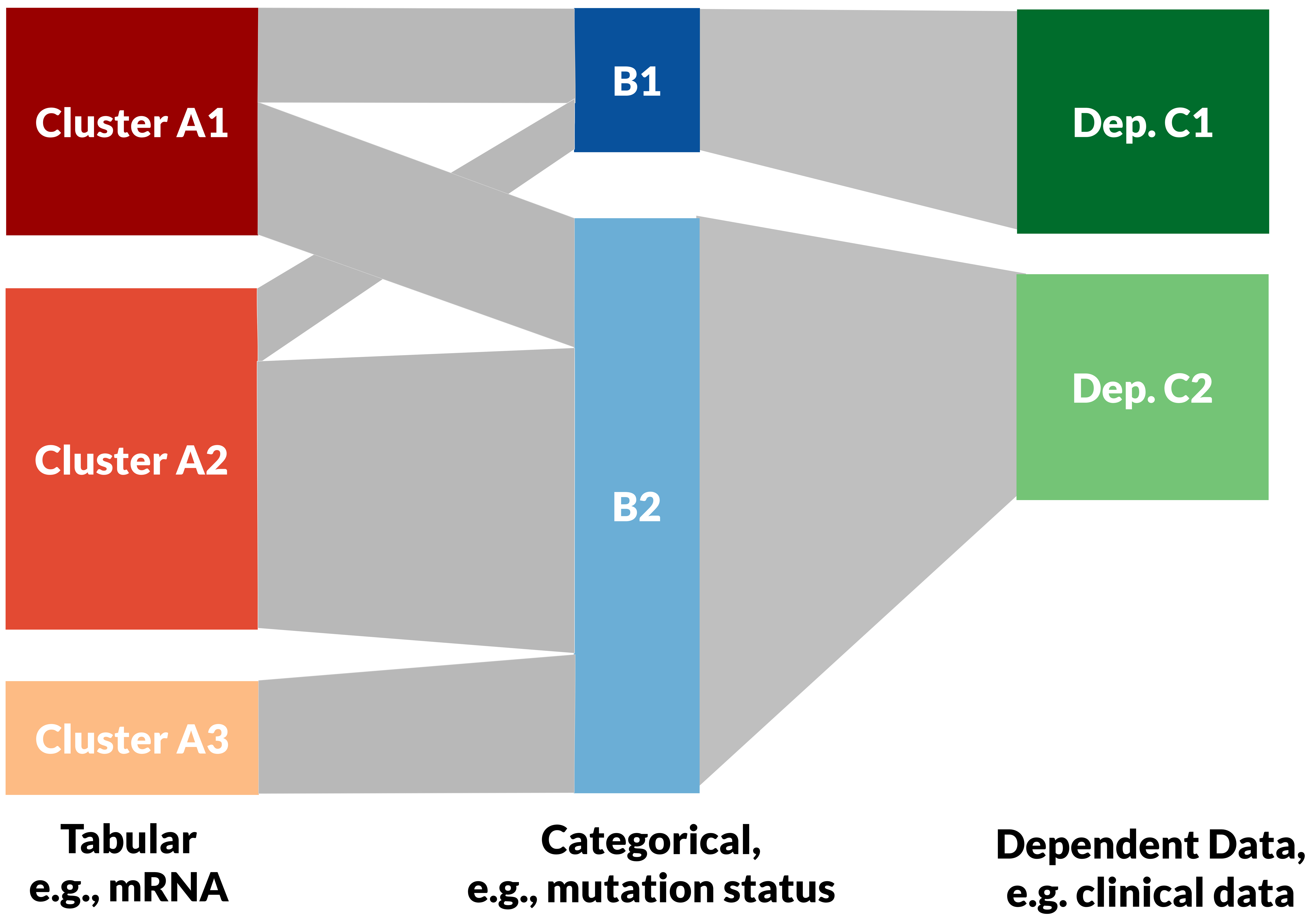
Clustering of mRNA Data

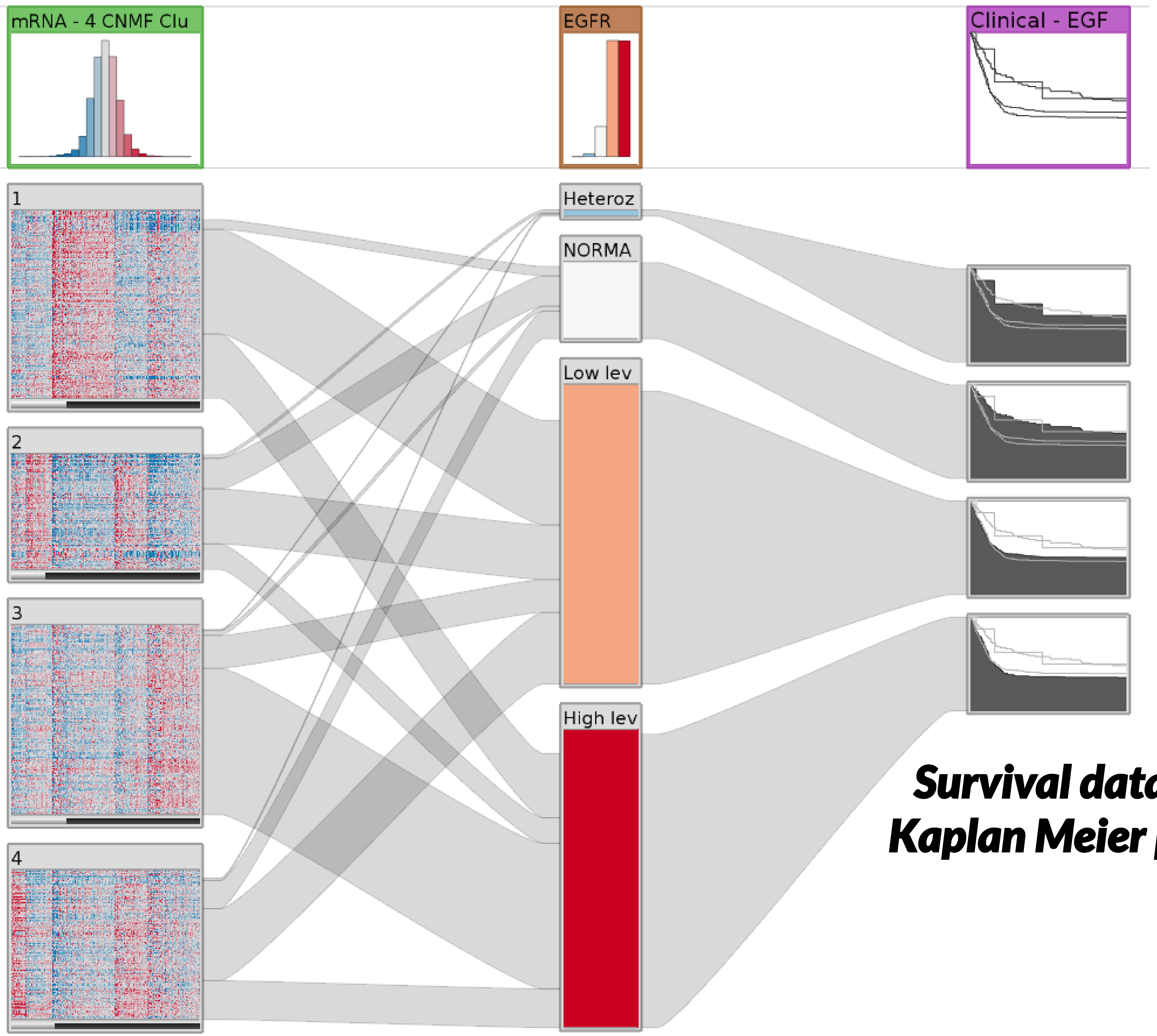
Stratification on Copy Number Status

Other Data – Same Stratification

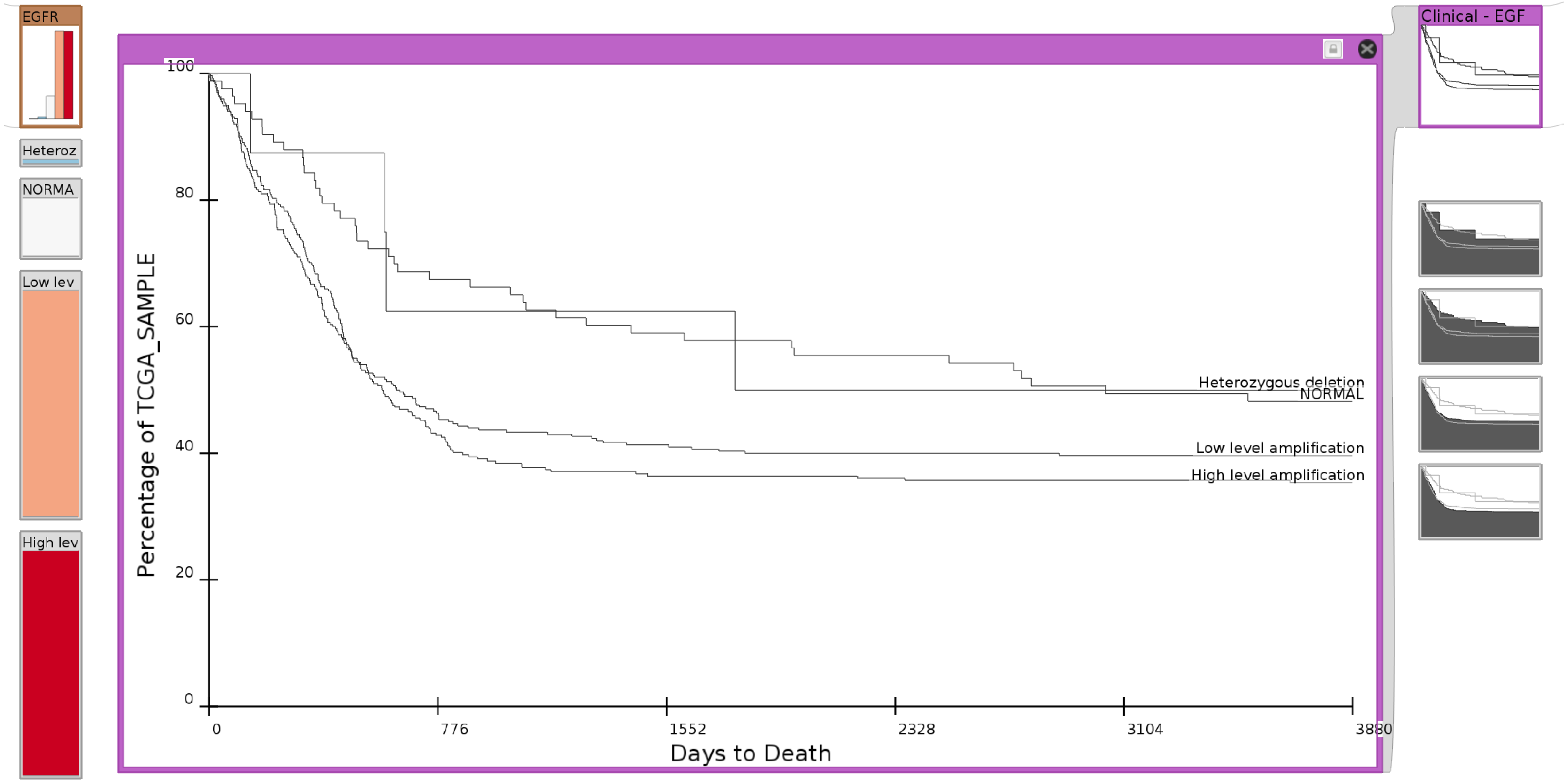


Other Data – Same Stratification

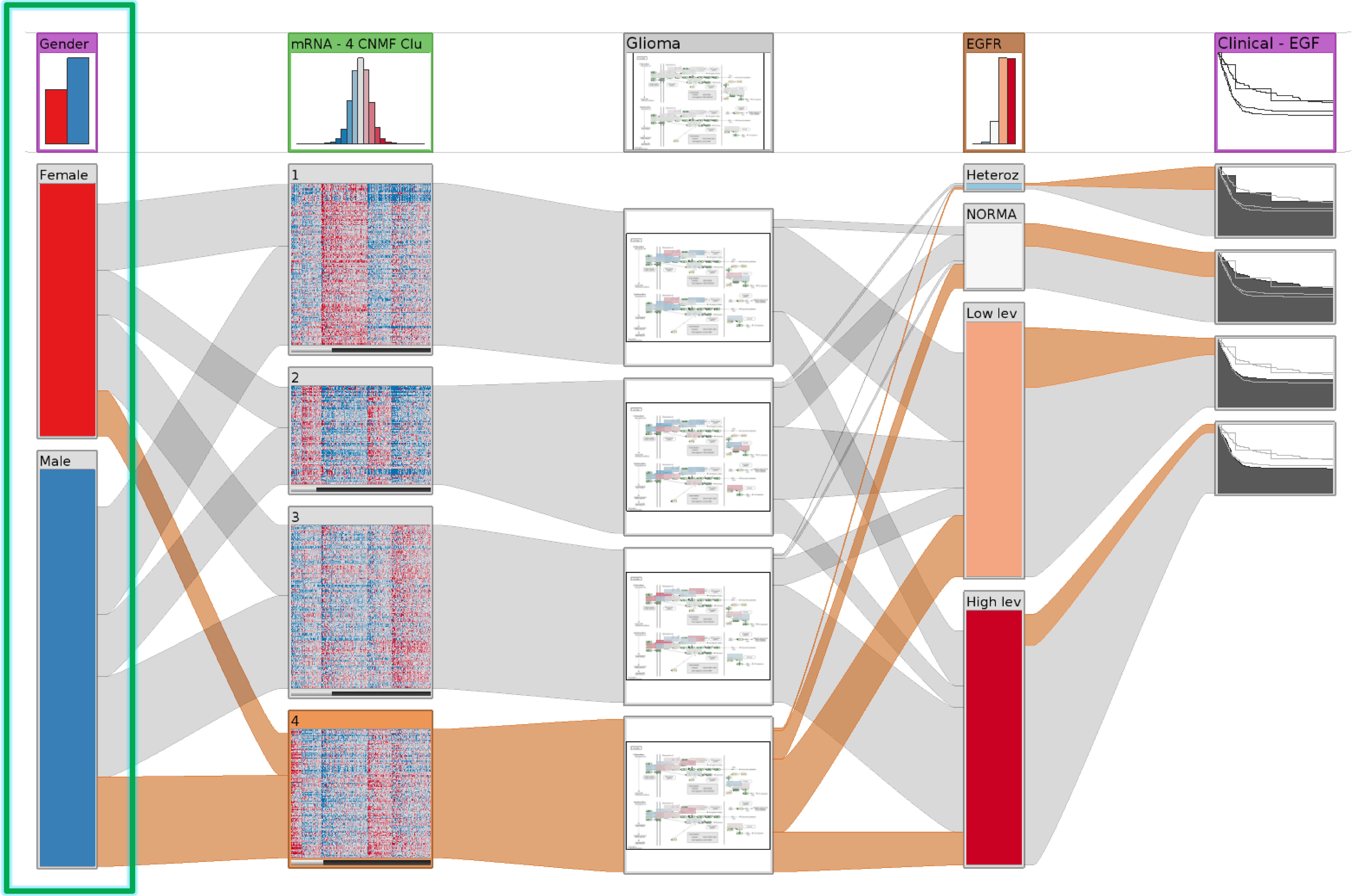




Survival data in Kaplan Meier plots


EGFR
Heteroz
NORMA
Low lev
High lev
Clinical - EGF
100
80
60
40
20
0
Percentage of TCGA_SAMPLE
0
776
1552
2328
3104
3880
Days to Death
Heterozygous deletion
NORMAL
Low level amplification
High level amplification

Detail View



Stratification based on clinical variable (gender)

Knowledge Driven Approach

⬇

Data Driven Approach

Finding Relevant Stratifications

~ 10 datasets

~ 15 clusterings per matrix

~ 15,000 stratifications for copy number & mutations

~ 500 pathways

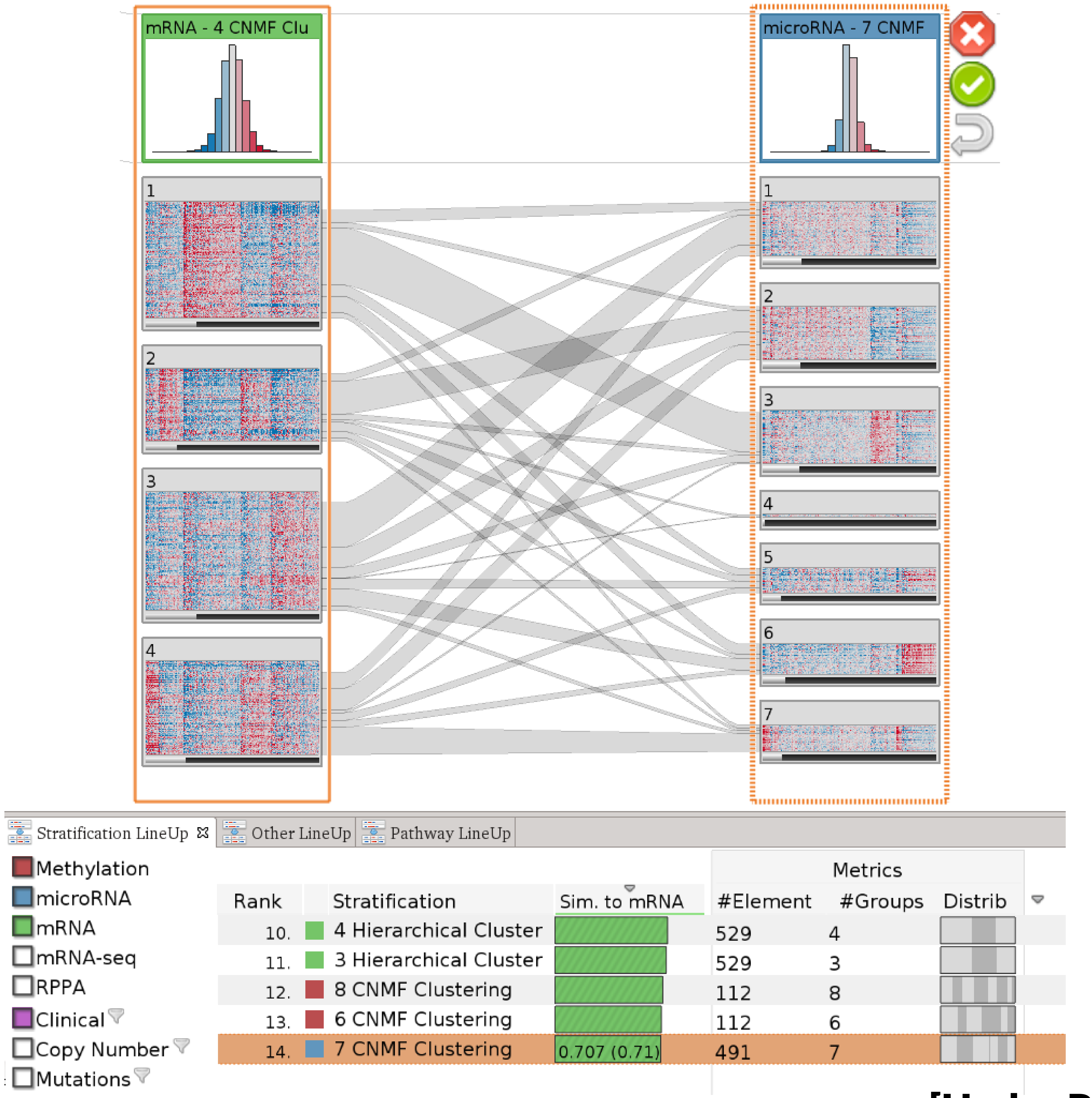
~ 20 clinical variables

Calculate scores for matches

Rank the results

Query column

Result column



Considered Datasets

Ranked Stratifications

[Under Review at Nature Methods]

Algorithms for finding..

... matching stratification

... matching subtype

... mutual exclusivity

... relevant pathway

... stratification with effect in survival

... high/low structural variation

Live-Demo!

http://stratomex.caleydo.org

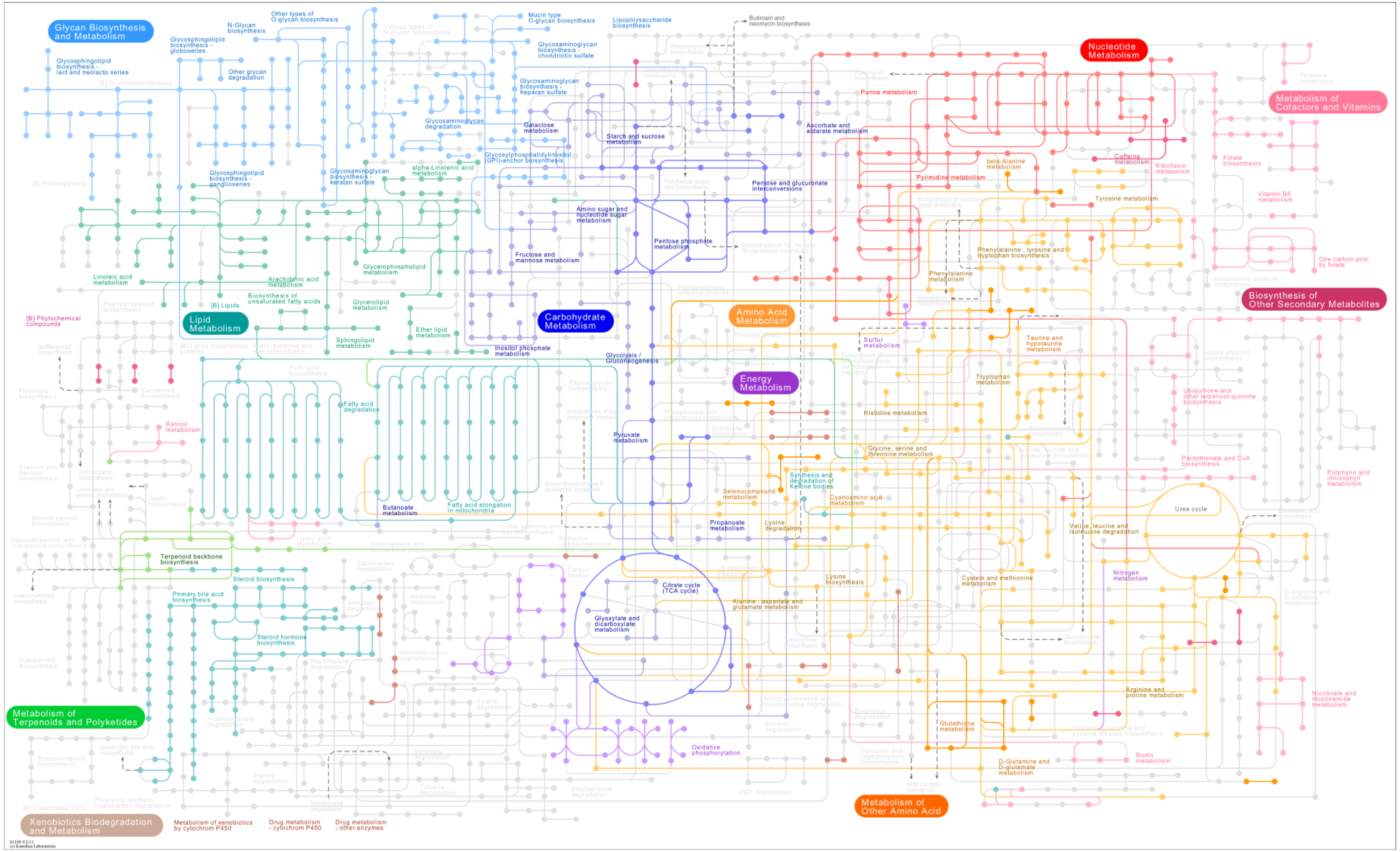
Current Developments: Generalization

[Under Review at IEEE TVCG (InfoVis)]

Pathways

Cytoskeleton remodeling
RacGEF
Rac
IKK
NFκB
Anti-apoptotic genes
DNA
PKB/Akt
Bad
Bcl-xl
Cell survival
Apoptosis
CASP9
Suppressed apoptosis
Raf
MEK
ERK
JNK
Proliferative genes
MAPK signaling pathway
Ral
RalBP1
PLD1
PA
Inhibition of apoptosis
VEGF signling pathway
STAT3
STAT1
VEGF
Angiogenesis
p16
CDK4/6
CyclinD1
Rb
Cell cycle
E2F
G1/S progression
K-Ras, HER2/neu
p16, p53, Smad4, BRCA2
p53 signaling pathway
DNA damage
p53
Karyotypic instability
Loss of growth inhibitory effects of TGFβ
TGF-β signaling pathway
BRCA2
Rad51
Failed repair of genes
DNA (dsDNA breaks)
BRAF
MAP2K1
MAPK1
CCND1
GRB2
SOS1
HRAS
ARAF
AR
KLK3
MYC
Glioma
CALML6
CAMK2A
MAPK signaling pathway
RAC1
MAP3K1
MAP2K4
FLNA
MAP3K8
Endometrial cancer
KRAS
ELK1
skin
lung
bre
RALGDS
RALA
RALBP1
Pancreatic cancer
MAPK signaling pathway

Background



Glycan Biosynthesis and Metabolism
Nucleotide Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Other Secondary Metabolites
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Energy Metabolism
Metabolism of Terpenoids and Polyketides
Xenobiotics Biodegradation and Metabolism
Metabolism of Other Amino Acid
Citrate cycle (TCA cycle)
Urea cycle
Oxidative phosphorylation
Purine metabolism
Pyrimidine metabolism
Glycolysis / Gluconeogenesis
Pentose phosphate metabolism
Fatty acid degradation
Steroid biosynthesis

Background


Whole Network
Pathway A
Pathway B
Pathway C
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q

Entourage

[Lex, InfoVis '13]

many pathways

pathway cross-connections

(large graphs)

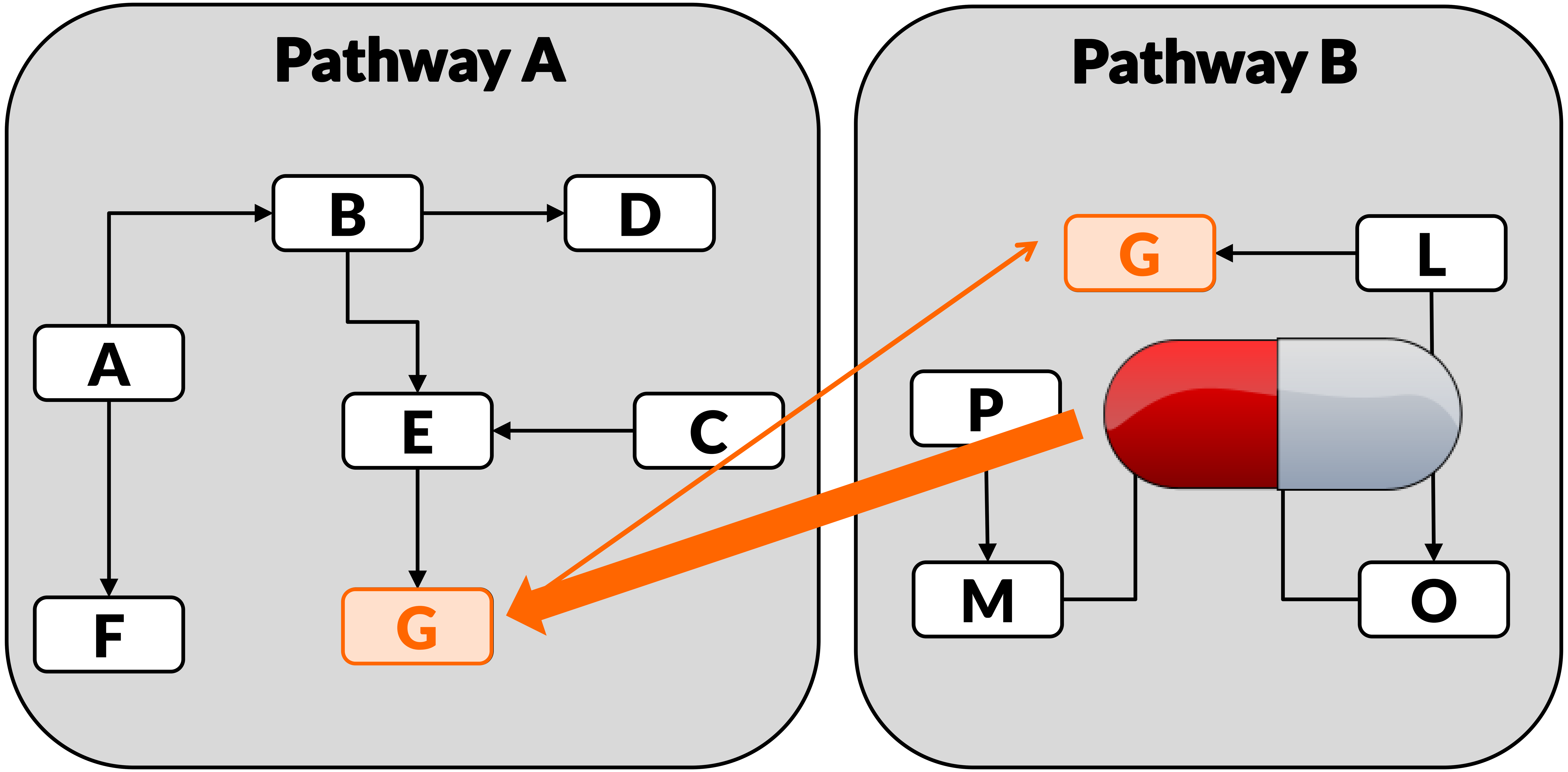
enRoute

[Partl, BioVis '12]
Best Paper Award
[Partl, BMC Bioinformatics '13]

experimental data on pathways

(multi-variate graphs)

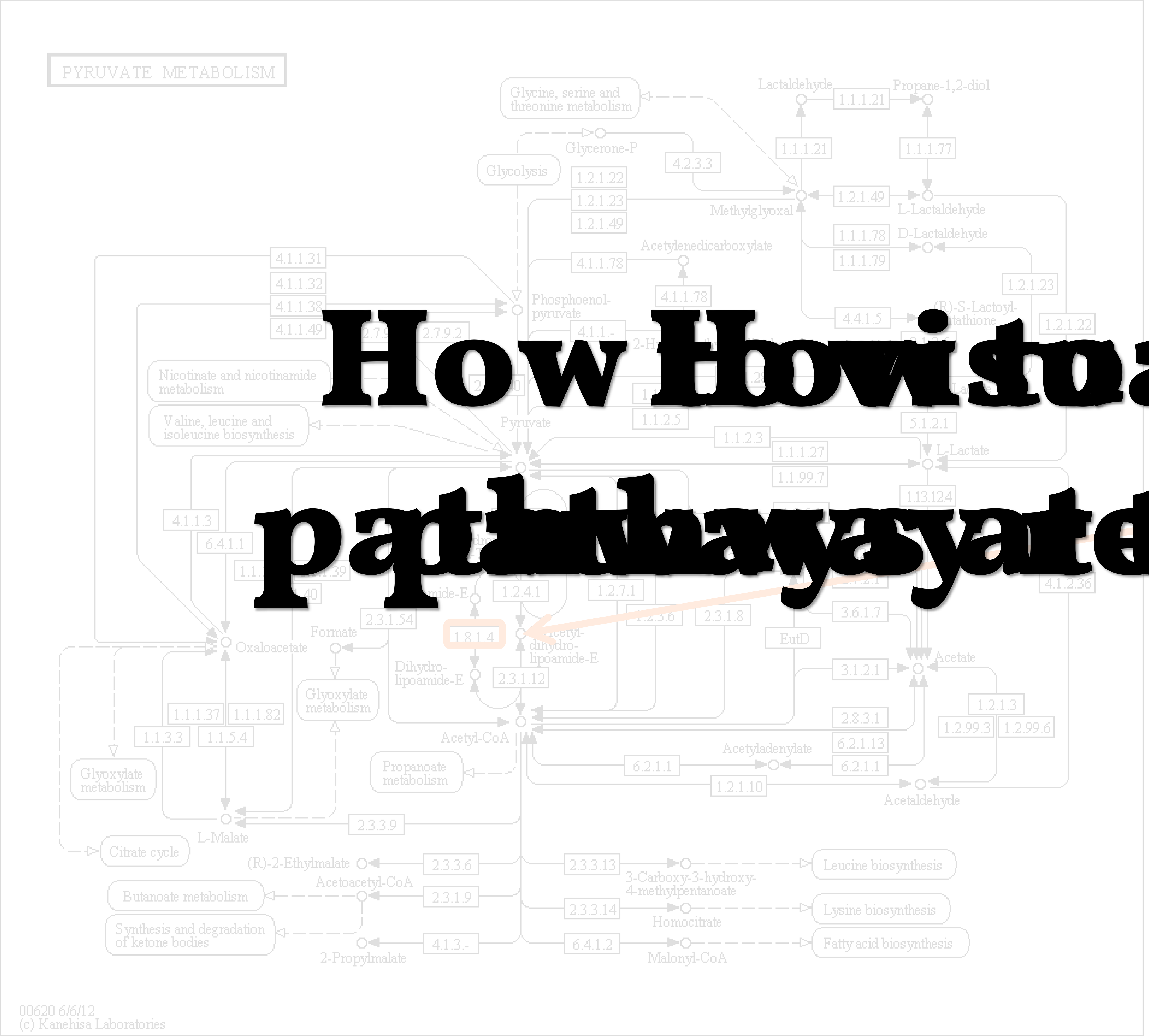
Challenges

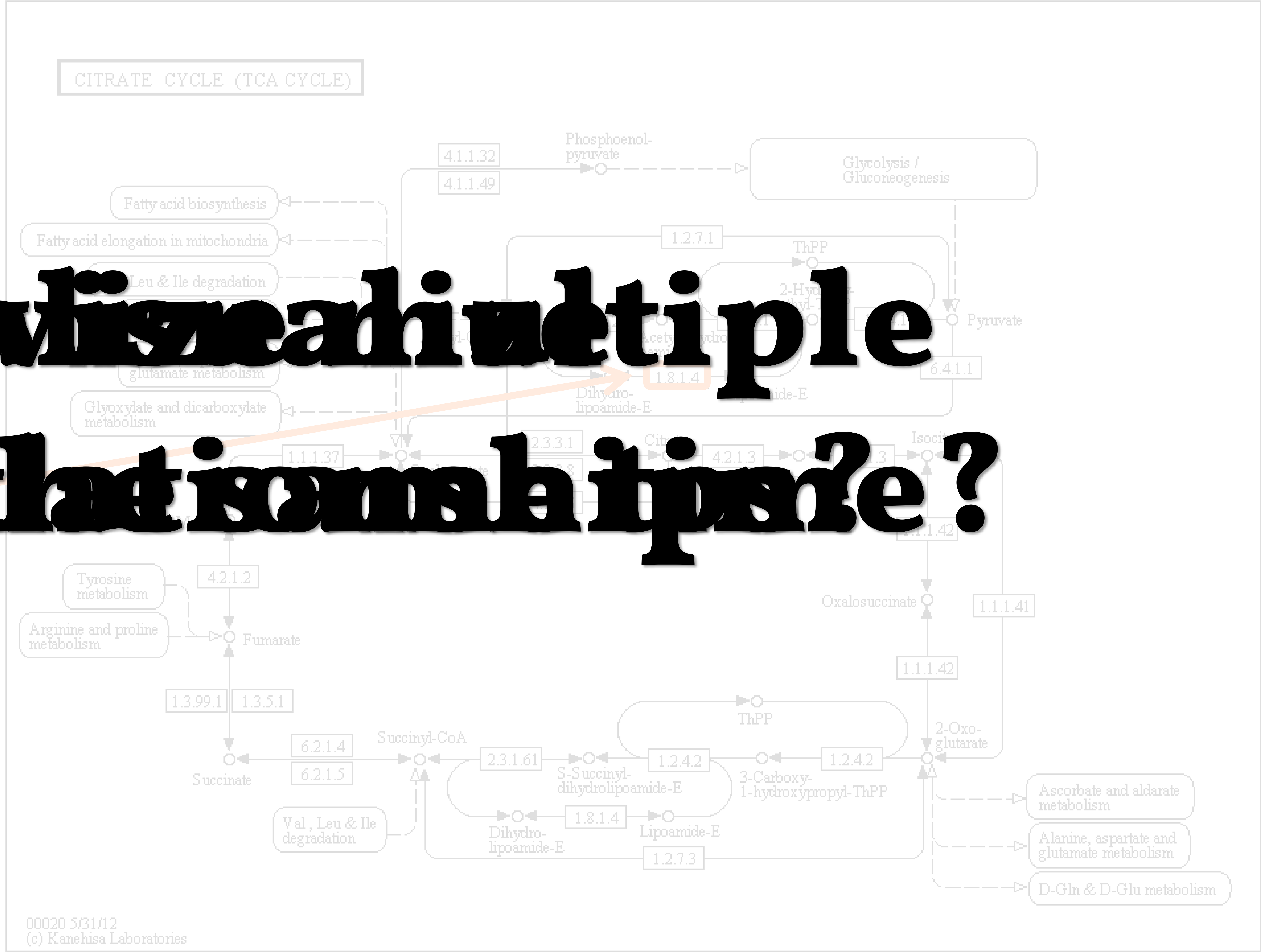


Drug side-effects

Drug repositioning

Challenges





How to visualize multiple pathways at the same time?

How to visualize the pathway relationships?

Challenges

How to visualize experimental data on pathways?

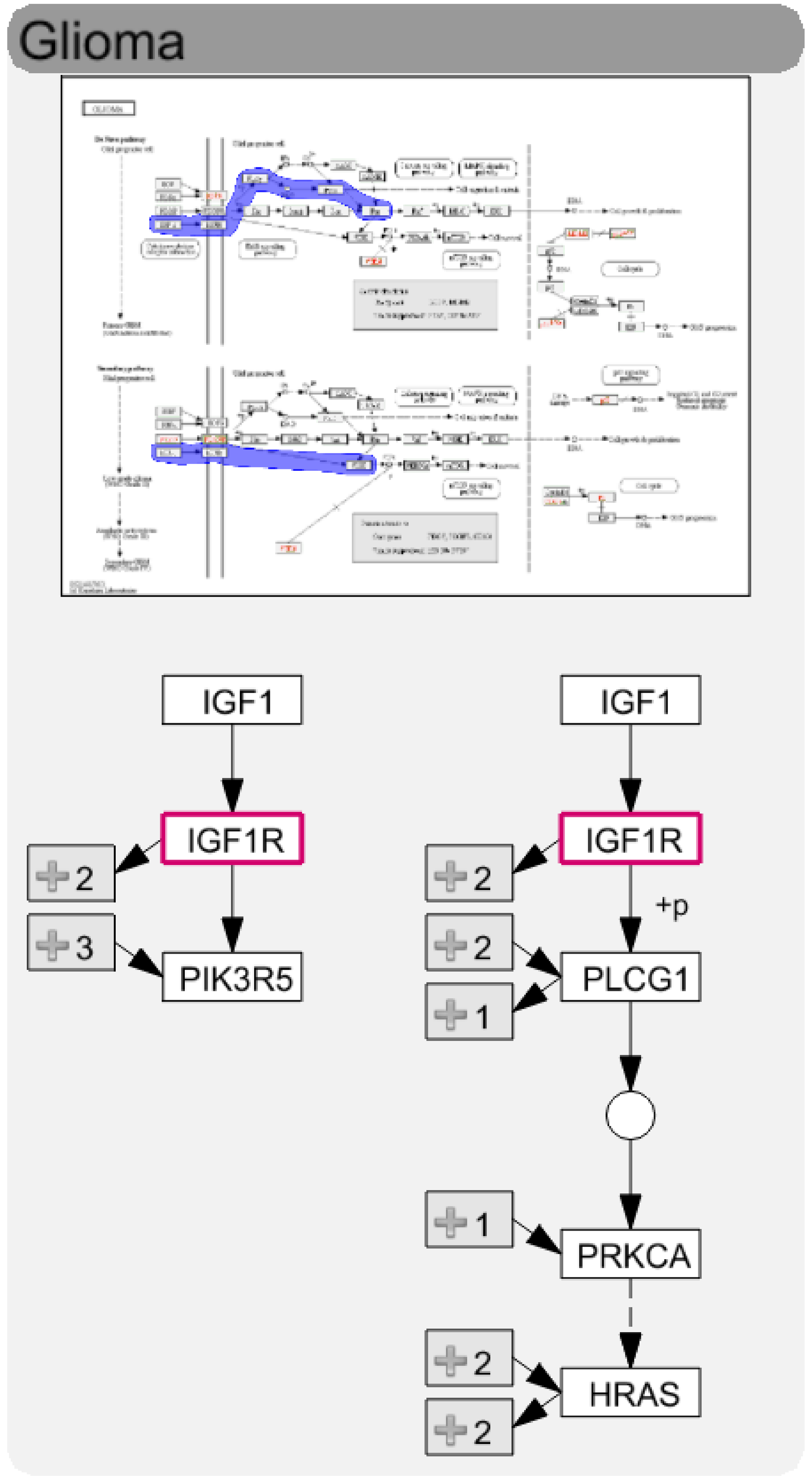
How to visualize multiple pathways at the same time?

How to visualize pathway relationships?

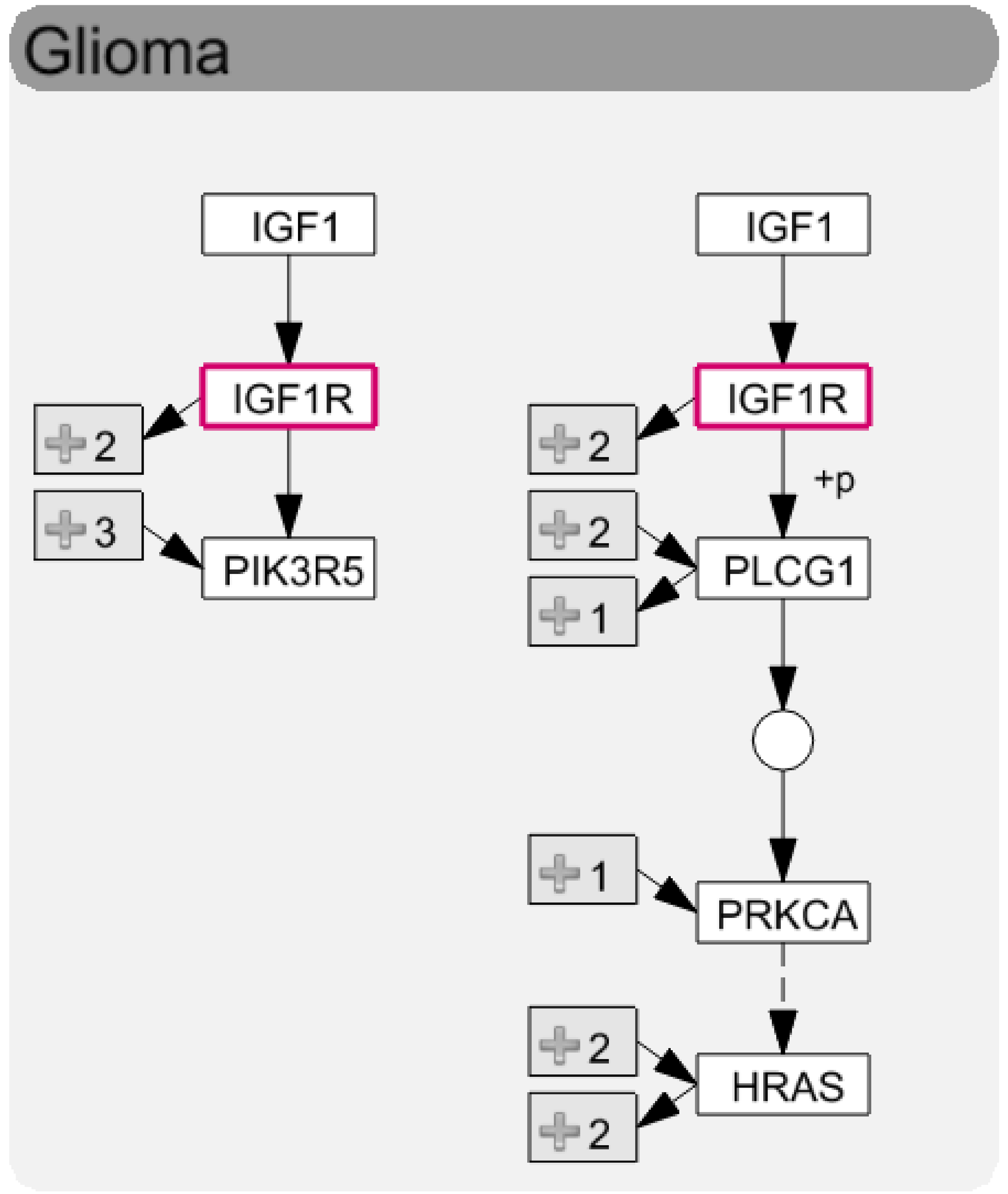
How to visualize experimental data on pathways?

Contextual Subsets

Levels of Detail



High



Medium

Low

Layout

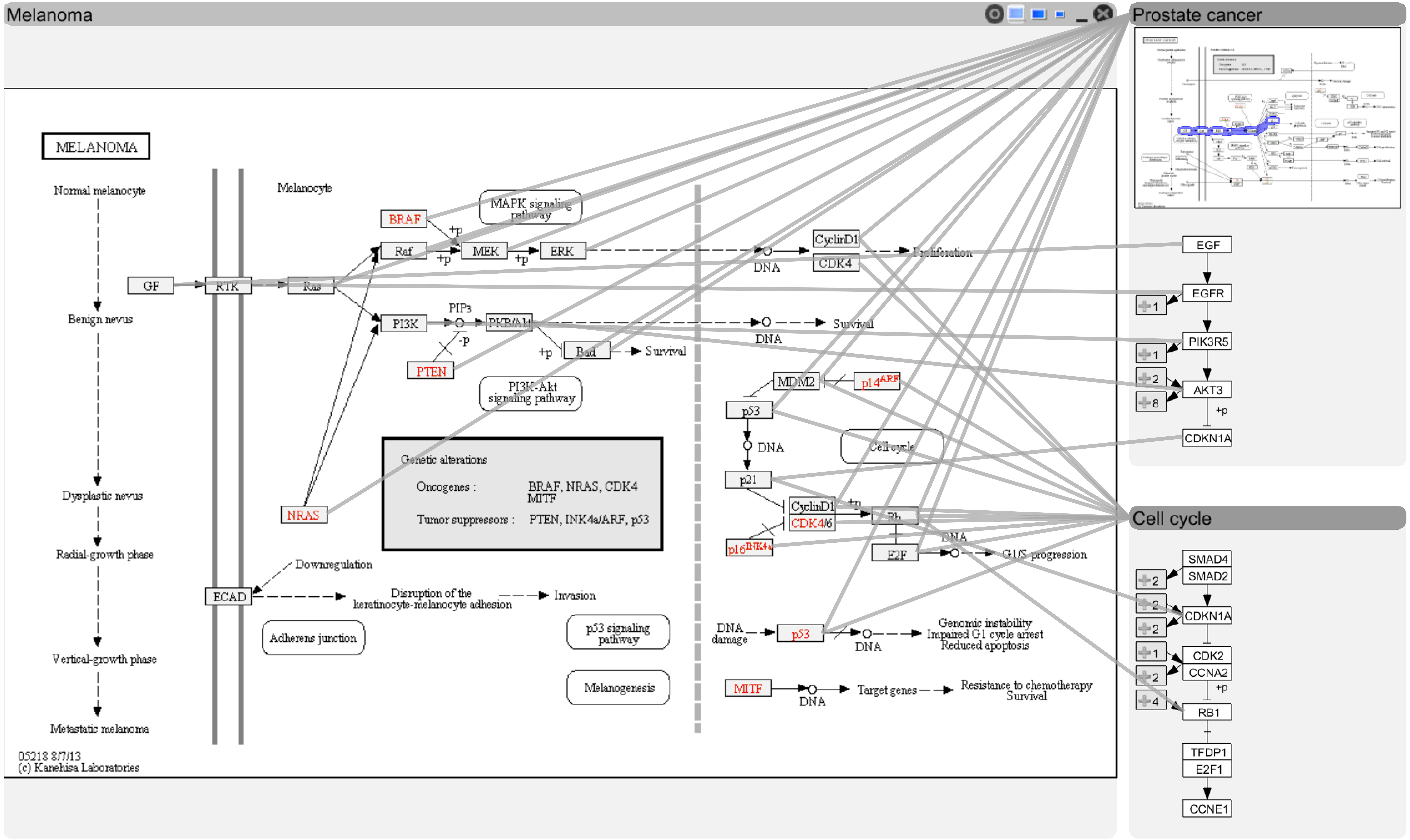

Glioma
IGF1
IGF1R
PIK3R5
PLCG1
+p
PRKCA
HRAS
Phospholipids
AA
PLA2
AA-CoA
ARTISt
PKA
cAMP
LPOX
EPOX
COX2
AA metabolites
DNA
CYP17
17β-HSD
3β-HSD
Pregnenolone
17-hydroxy-pregnenolone
Dehydroepi-androstenedione
Androstenediol
Progesterone
17-hydroxy-progesterone
Androstenedione
Testosterone
Circulation
CYP1B1
4-OH-estradiol
CYP19A1
17β-estradiol
CYP1A1
2-OH-estradiol
Estrone
Follicular fluid
04913 4/22/13
(c) Kanehisa Laboratories
Prostate cancer
EGF
EGFR
AKT3
CDKN1A
Proteoglycans in cancer
GPC3
IGF2
MAPK1
DCN
PDK1
RPS6KB
EIF4B
RPS6
PI3K-Akt signaling pathway
CSF1
CSF1R
IRS1
PIK3CA
HIF-1 signaling pathway
MTOR
EIF4EBP
Progesterone-mediated oocyte
KRAS
ARAF
Oocyte meiosis
Melanoma

How to visualize multiple pathways
at the same time?

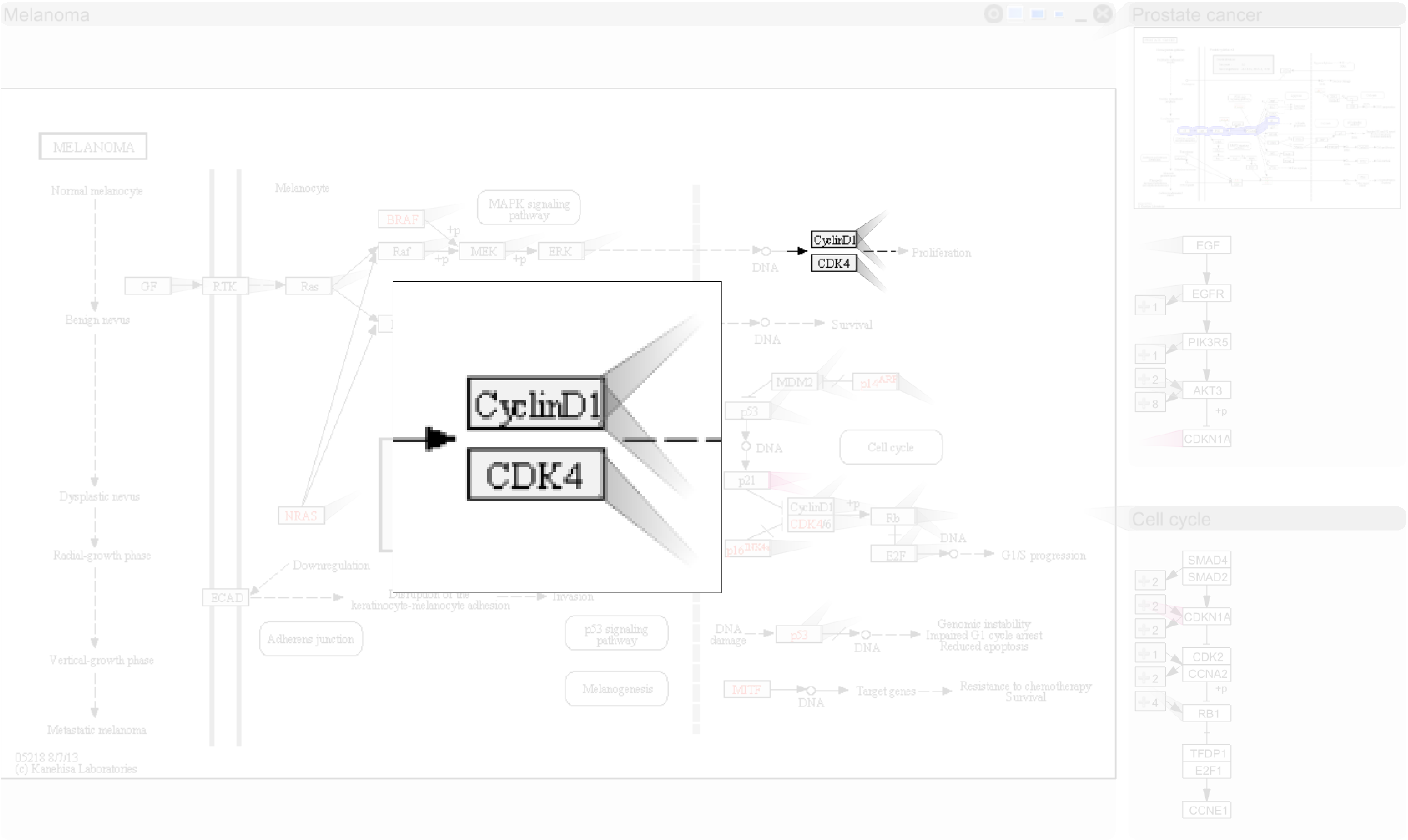
How to visualize
pathway relationships?

How to visualize
experimental data on pathways?

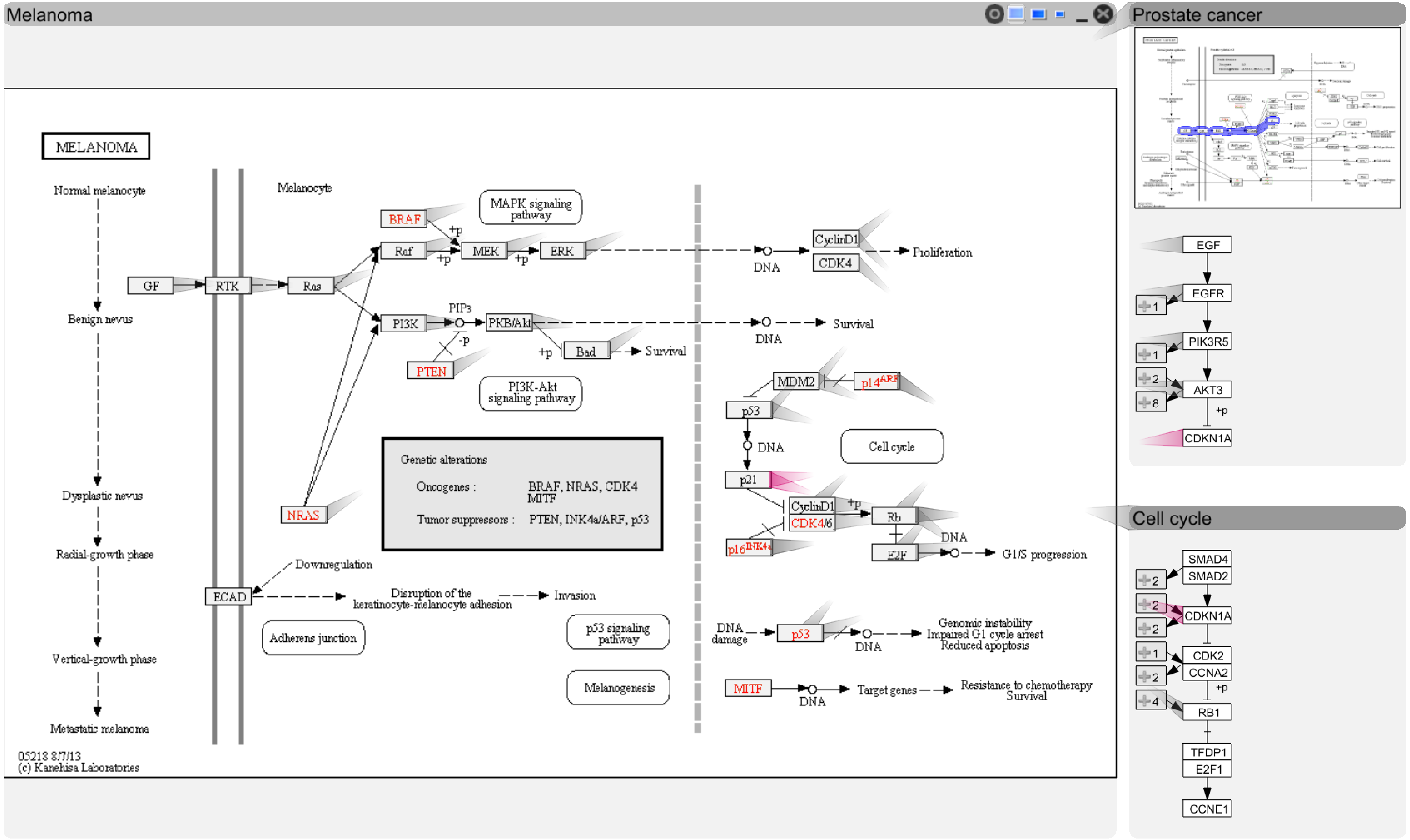
Visualizing Relationships



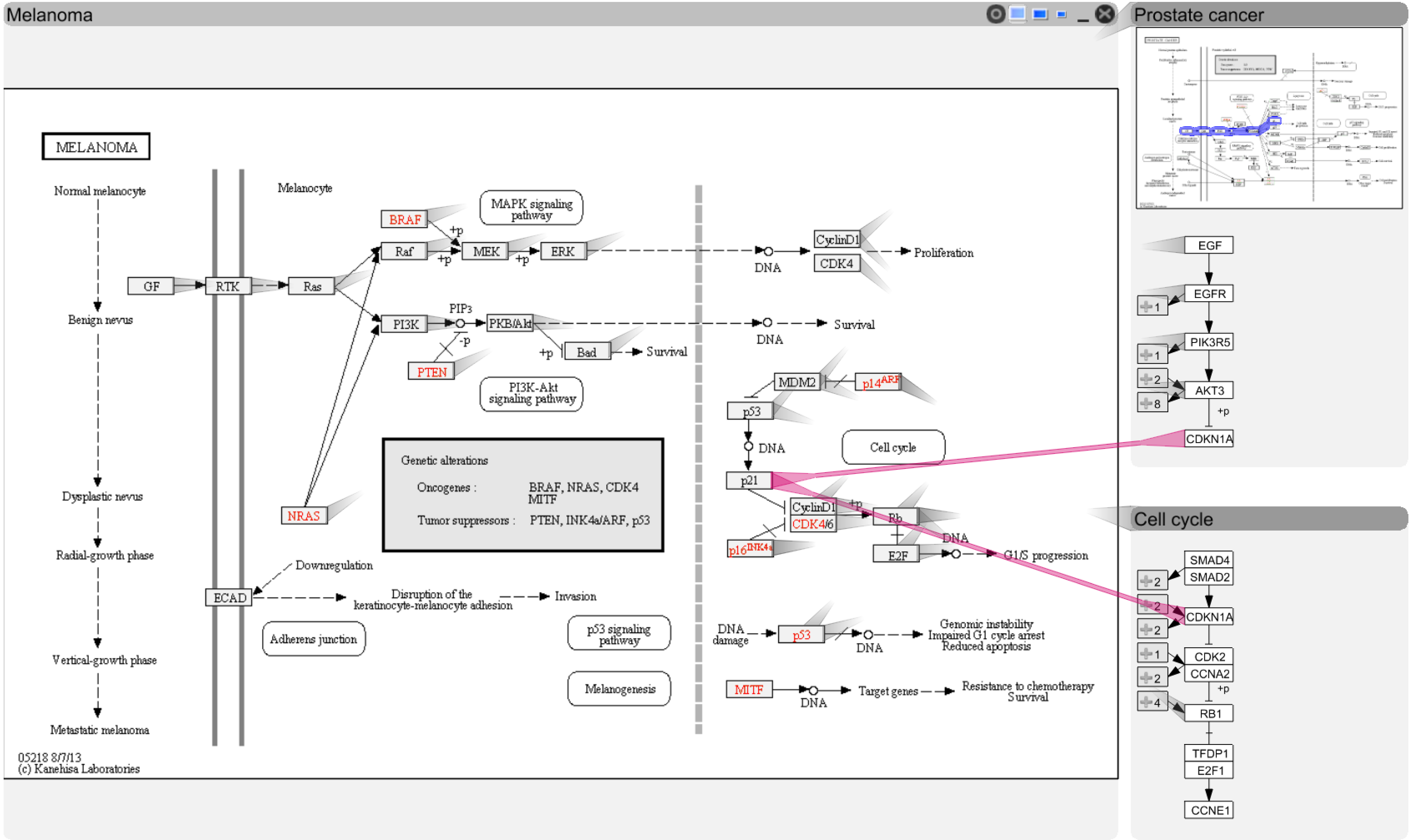
Visualizing Relationships



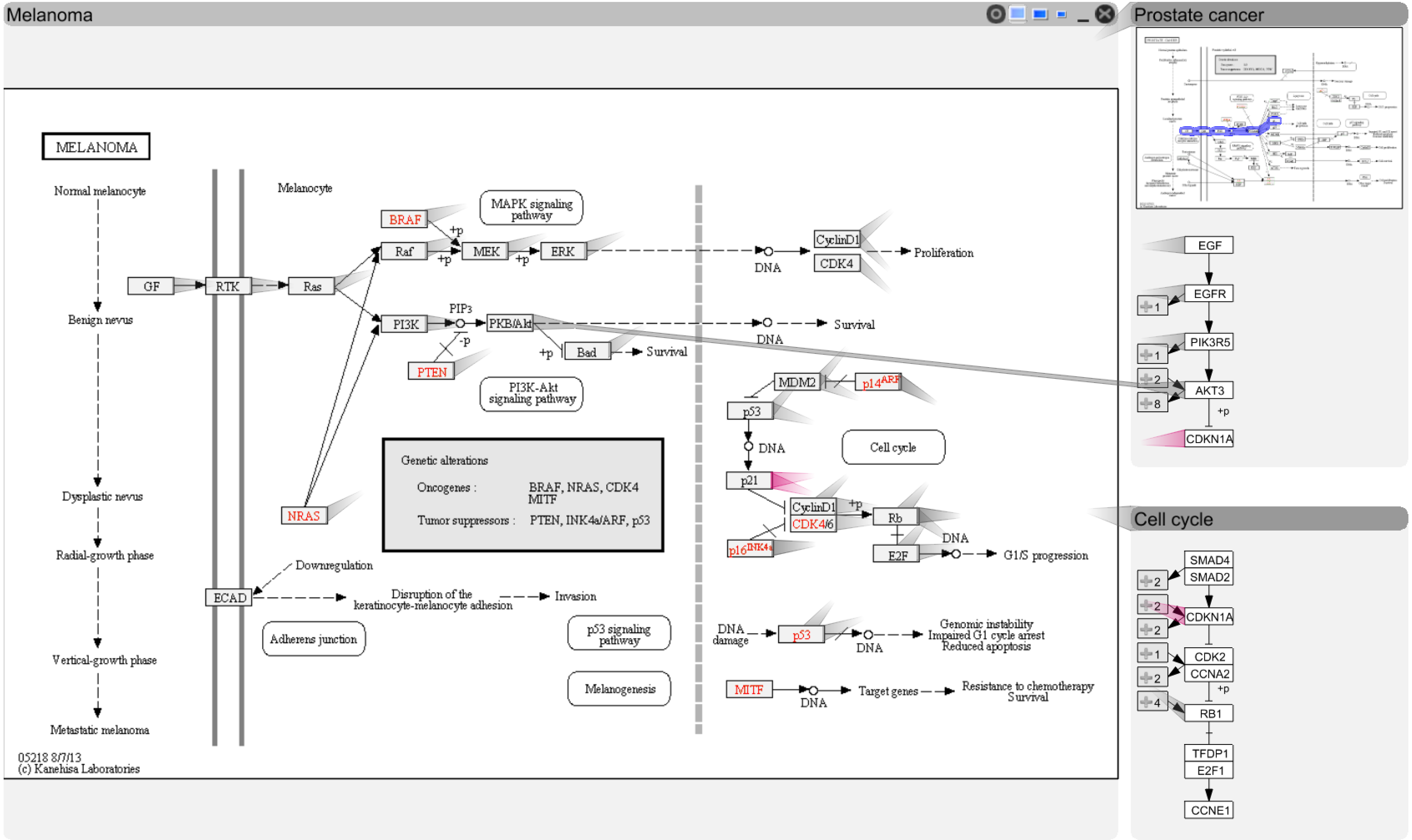
Visualizing Relationships



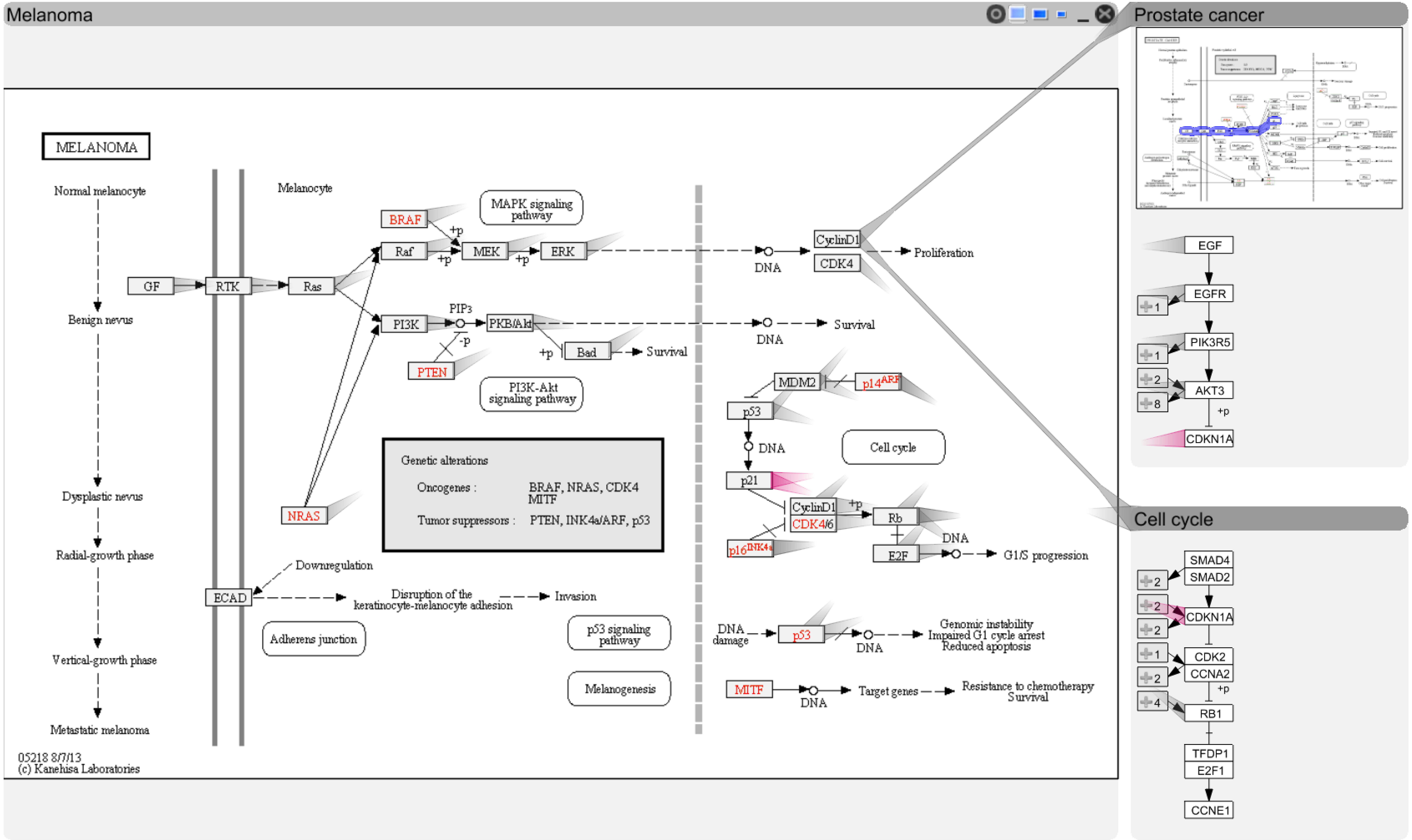
Visualizing Relationships



Visualizing Relationships



Visualizing Relationships



How to visualize multiple pathways
at the same time?

How to visualize
pathway relationships?

How to visualize
experimental data on pathways?

Experi-mental Data and Pathways

enRoute

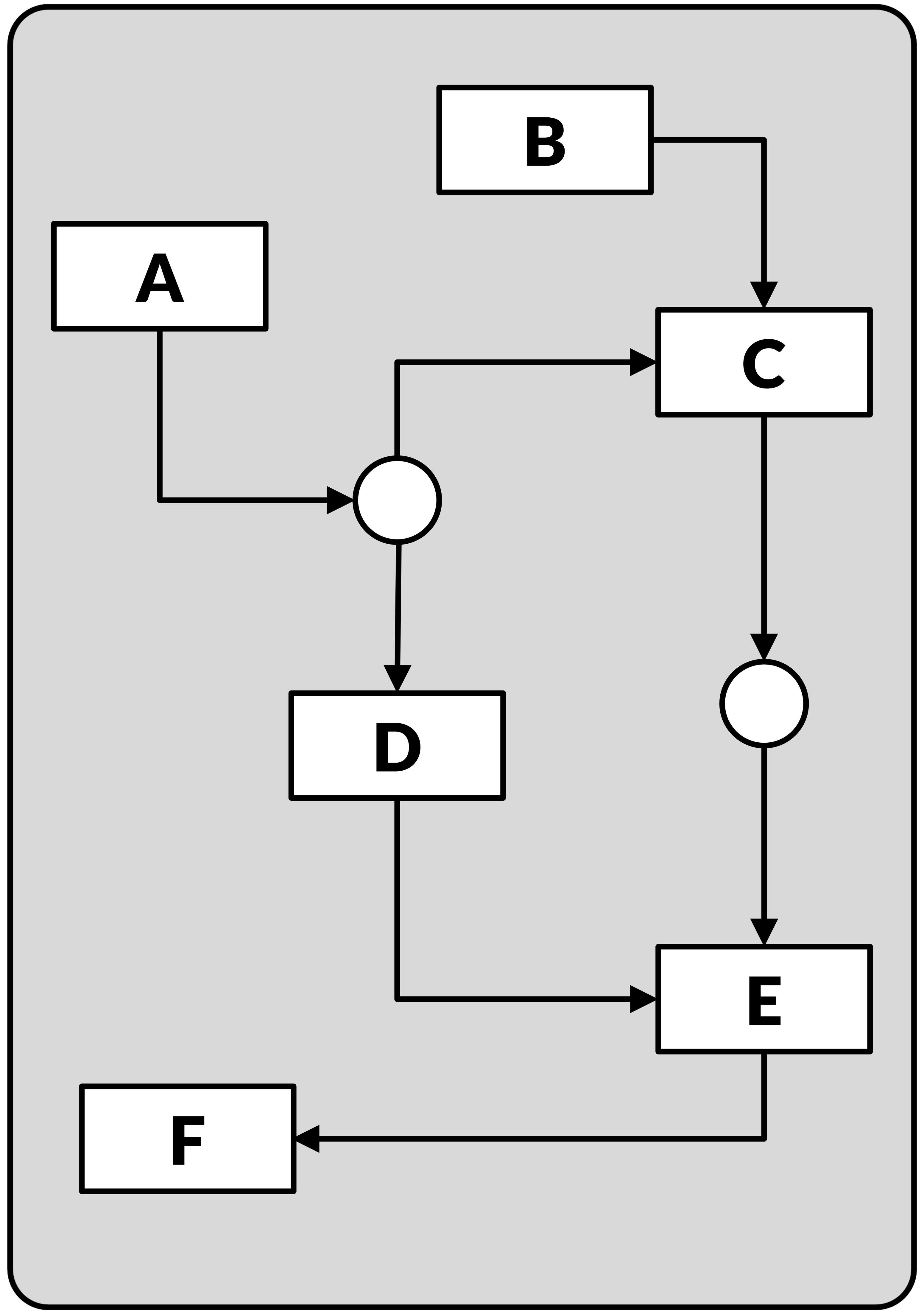
[Partl, BioVis '12]

Cannot account for variation found in real-world data

Branches can be (in)activated due to

- mutation,
- changed gene expression,
- modulation due to drug treatment,
- etc.

Good Old Color Coding



A	-3.4
B	2.8
C	3.1
D	-3
E	0.5
F	0.3

Challenge: Data Scale & Heterogeneity

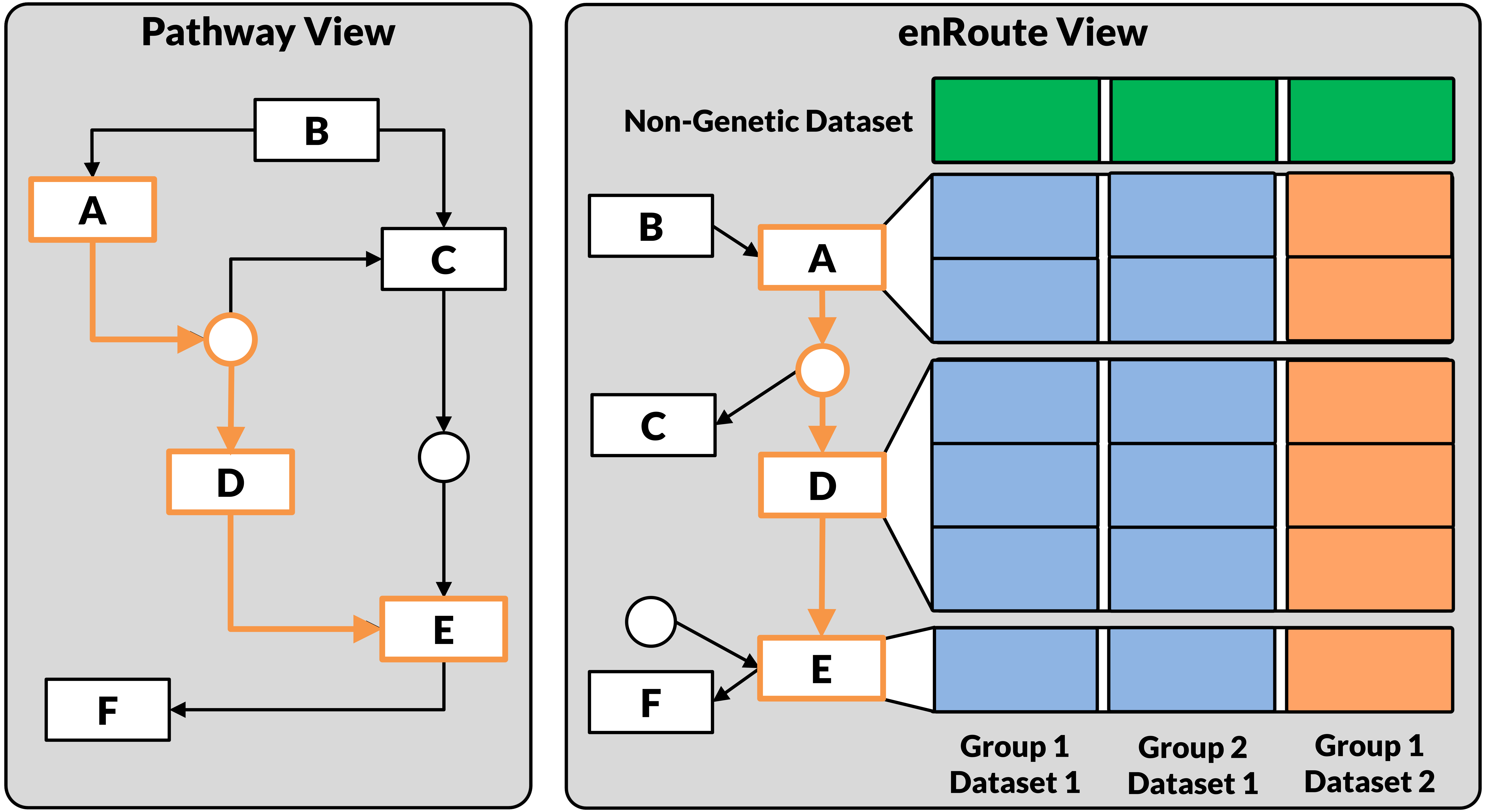
Large number of experiments

- Large datasets have more than 500 experiments

Multiple groups/conditions

Different types of data, require different visualization techniques

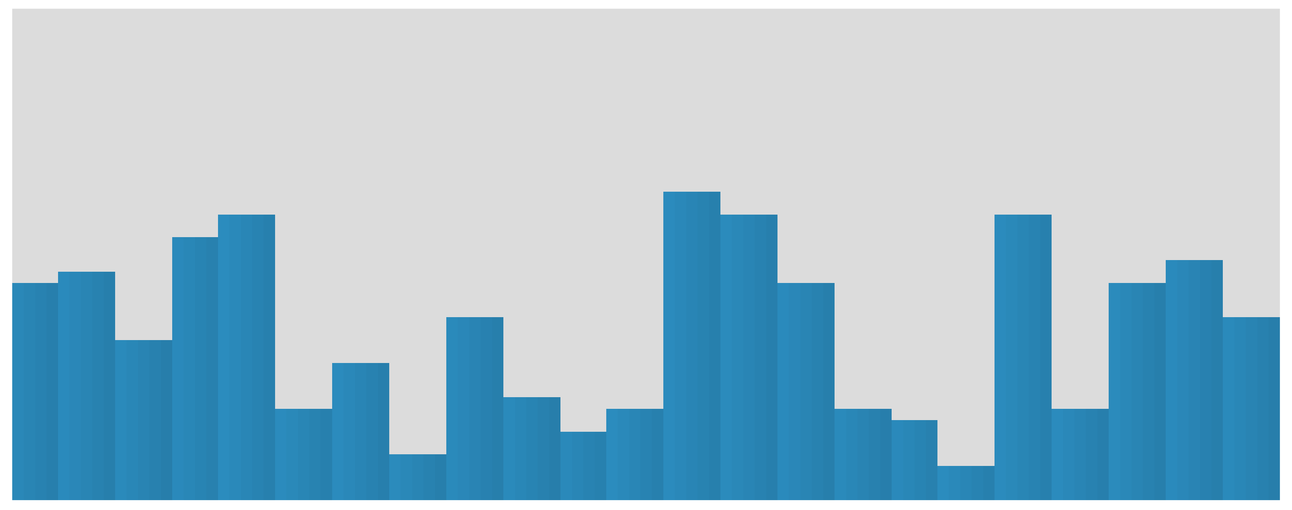
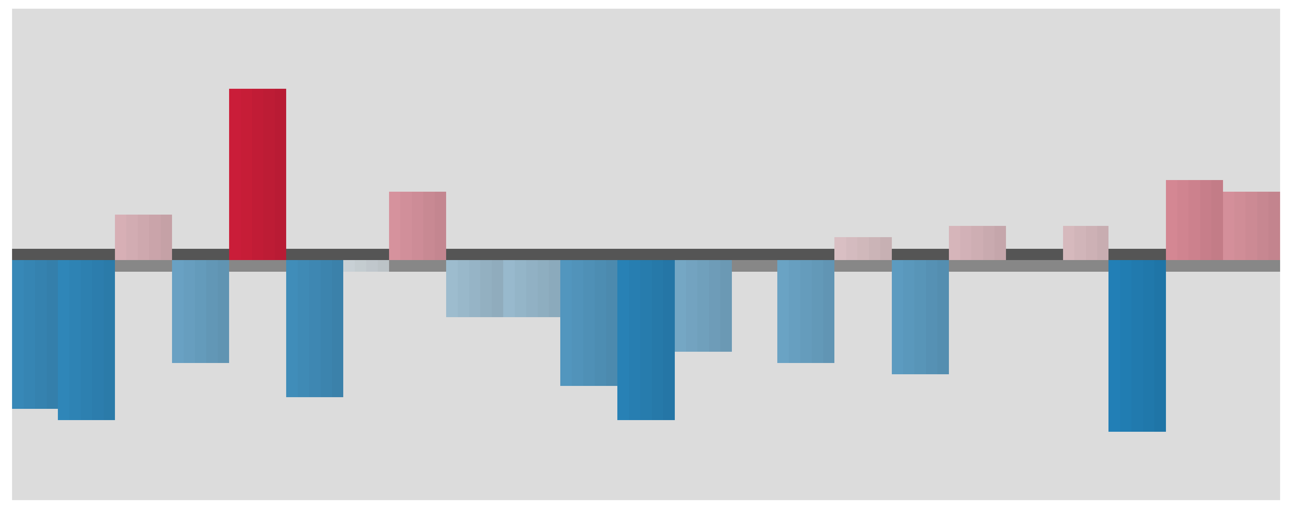
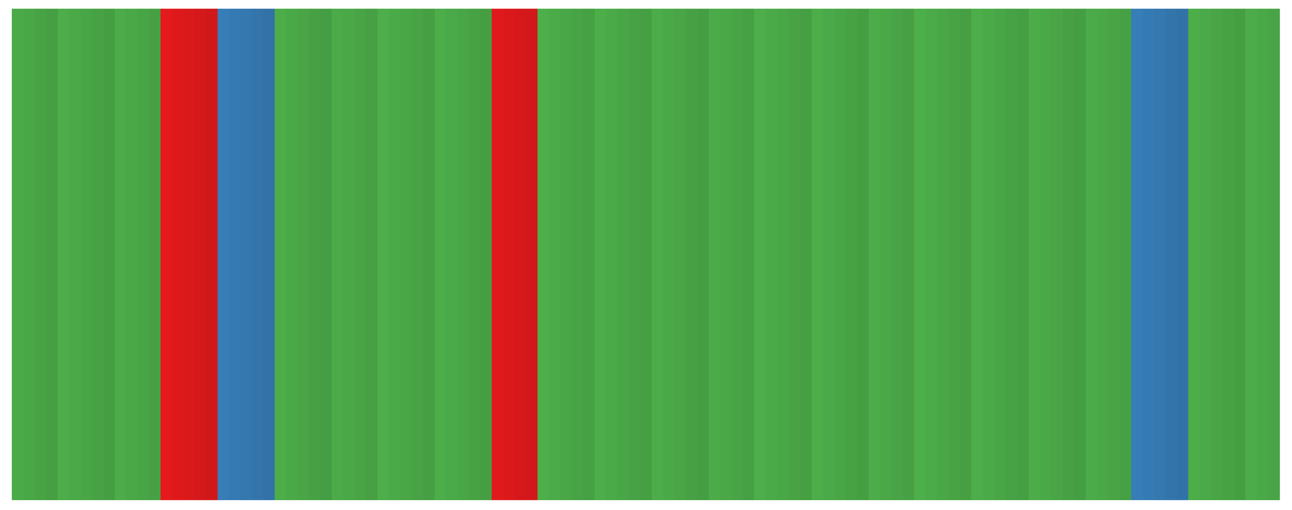
Concept



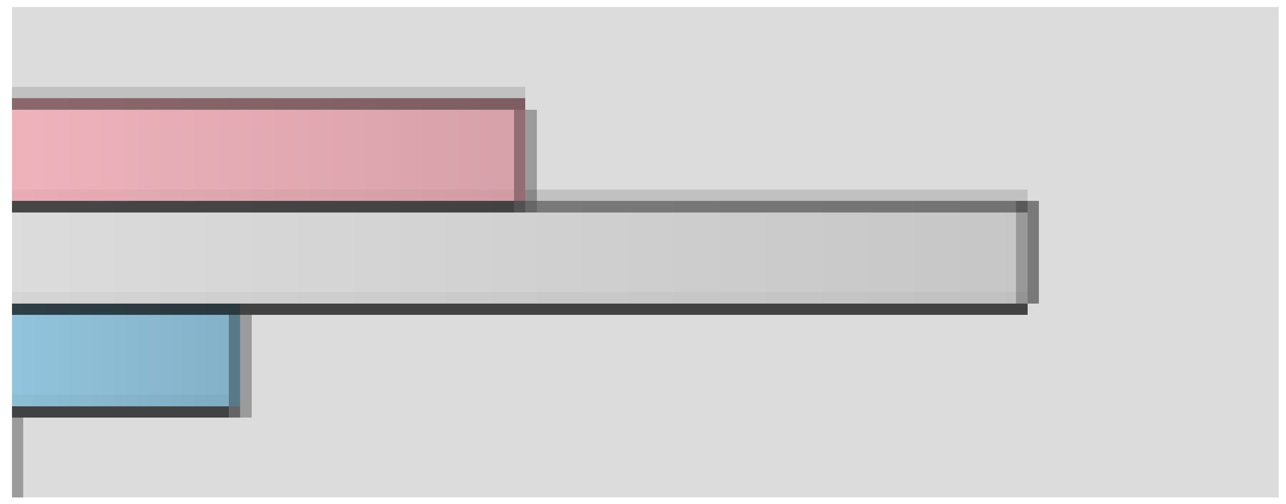
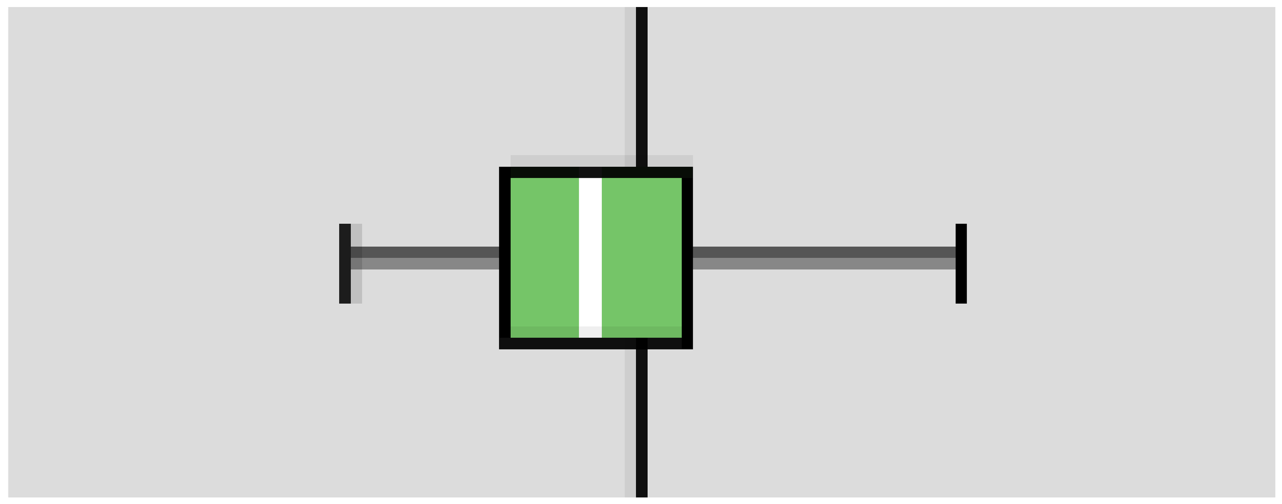
Pathway View
B
A
C
D
E
F
enRoute View
Non-Genetic Dataset
B
A
C
D
E
F
Group 1
Dataset 1
Group 2
Dataset 1
Group 1
Dataset 2

Experimental Data Representation

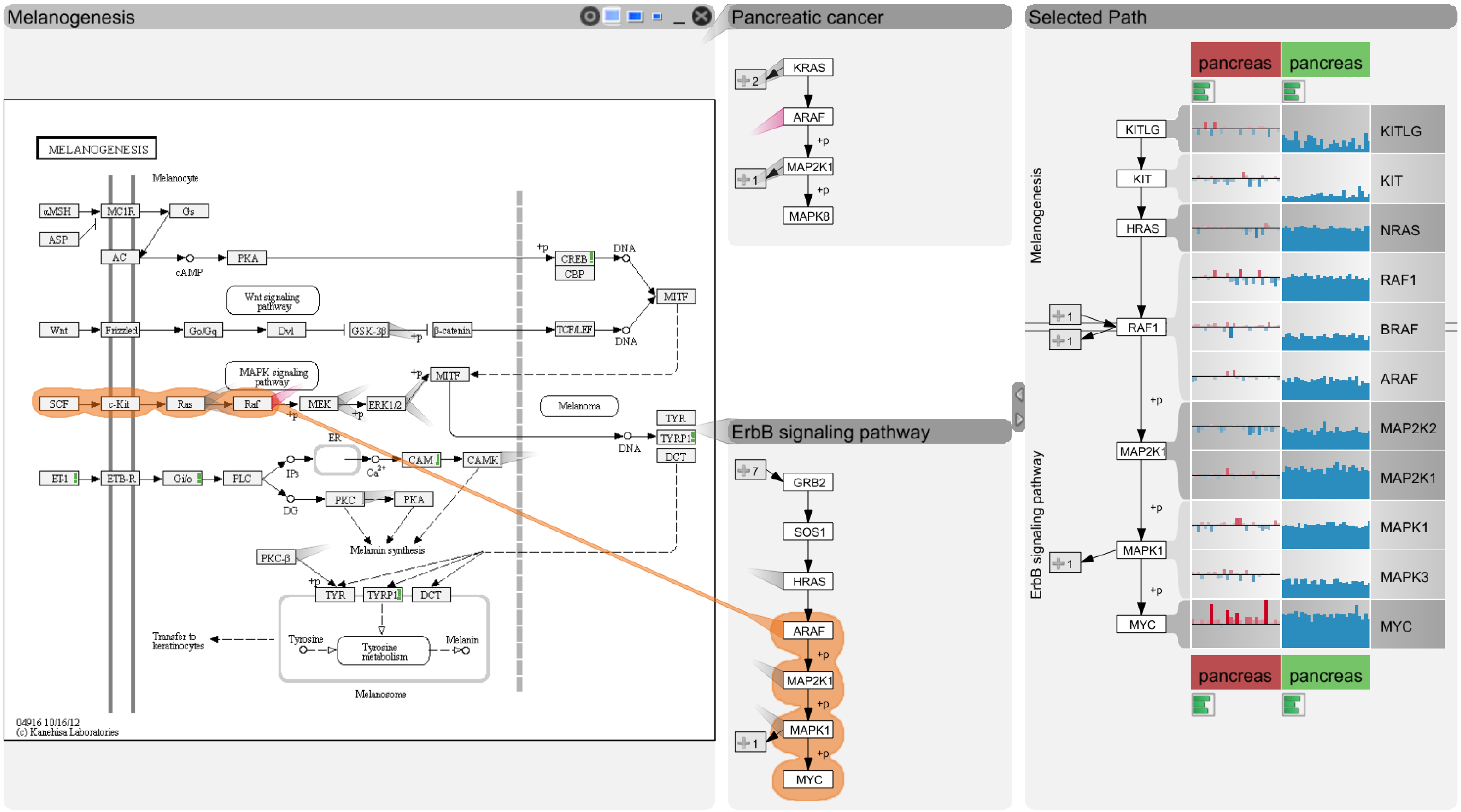
Individual Dimensions

Summary Representations

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
CAM
CAMK
ET-1
ETB-R
Gi/o
PLC
IP3
Ca2+
DG
PKC
PKC-β
Melanin synthesis
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
MAPK1
MYC
Selected Path
pancreas
pancreas
Melanogenesis
ErbB signaling pathway
KITLG
KIT
HRAS
NRAS
RAF1
BRAF
ARAF
MAP2K2
MAP2K1
MAPK1
MAPK3
MYC
pancreas
pancreas

Live-Demo!

http://entourage.caleydo.org
http://enroute.caleydo.org

Rankings

[Gratzl, InfoVis '13]

Best Paper Award

Rankings are omnipresent

Goal

Intuitive
Interactive
Multi-Attribute
Ranking Visualization
To Create
Refine
Explore

10 Requirements

University
MIT, USA
Harvard, USA
Princeton, USA
Cambridge, UK
Oxford, UK

10 Requirements

University

Encode Rank

Princeton, USA

Cambridge, UK

Oxford, UK

10 Requirements

1. Encode Rank

Encode Rank

Rank	University
1.	MIT, USA
2.	Harvard, USA
3.	Princeton, USA
4.	Cambridge, UK
5.	Oxford, UK

10 Requirements

1. Encode Rank

Encode Cause of Rank

10 Requirements

1. Encode Rank
2. Encode Cause of Rank

Encode Cause of Rank

Rank	University	Score
1.	MIT, USA	
2.	Harvard, U	
3.	Princeton,	
4.	Cambridge	
5.	Oxford, UK	4.0

10 Requirements

1. Encode Rank
2. Encode Cause of Rank

Support Multiple Attributes

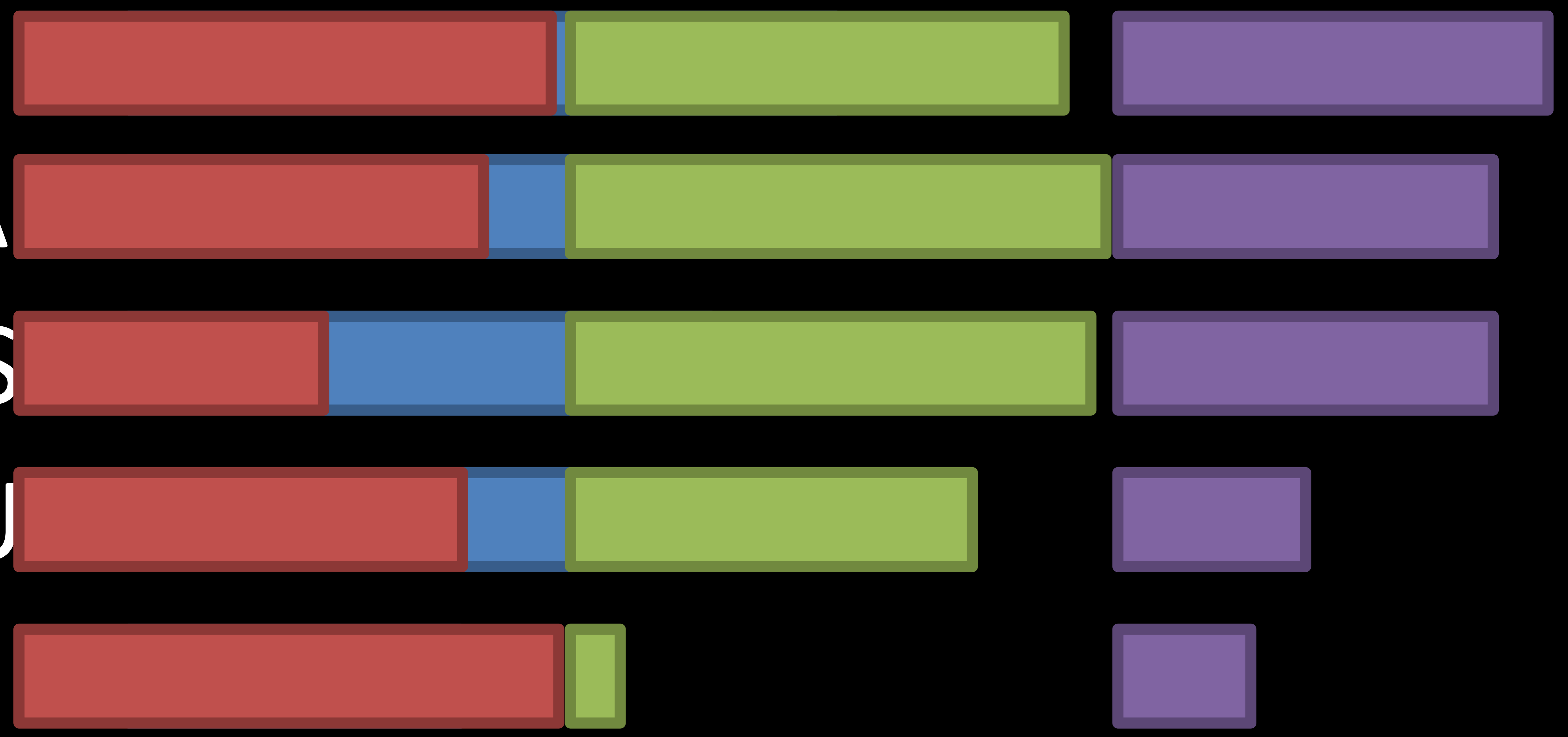
10 Requirements

1. Encode Rank
2. Encode Cause of Rank
3. Support Multiple Attributes

Support Multiple Attributes

Score = f(A, B, C)

Rank	University	A	B	C
1.	MIT, USA			
2.	Harvard, USA			
3.	Princeton, US			
4.	Cambridge, U			
5.	Oxford, UK			



Combiner functions: f(A,B,C)

(Weighted) sum

$Score = w_a A + w_b B + w_c C$

Maximum

$Score = max(A, B, C)$

Product

Nesting

…

→Serial

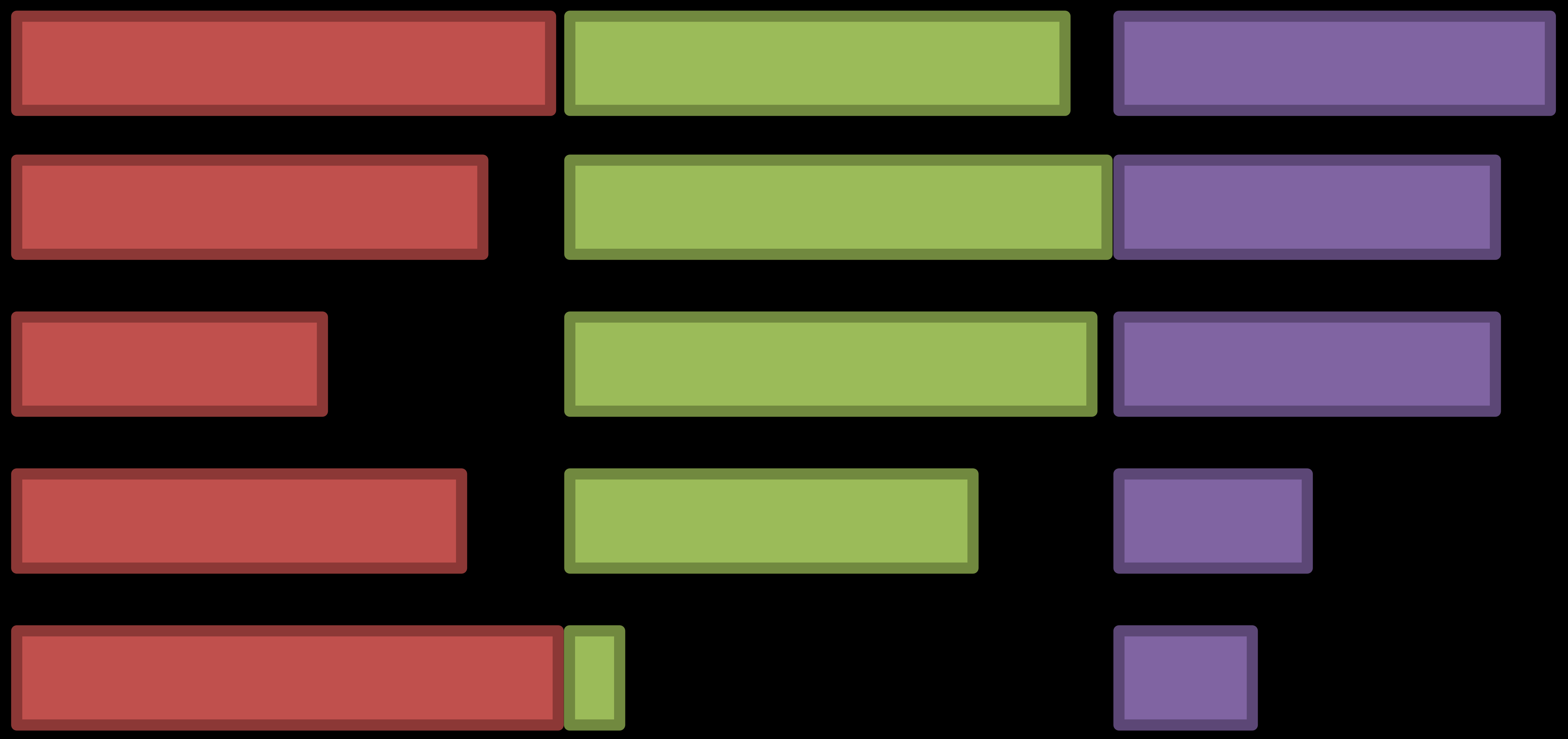
→Parallel

→Complex Combiners

Serial Combiner (as Stacked Bar)

$w_a A \quad + \quad w_b B \quad + \quad w_c C$

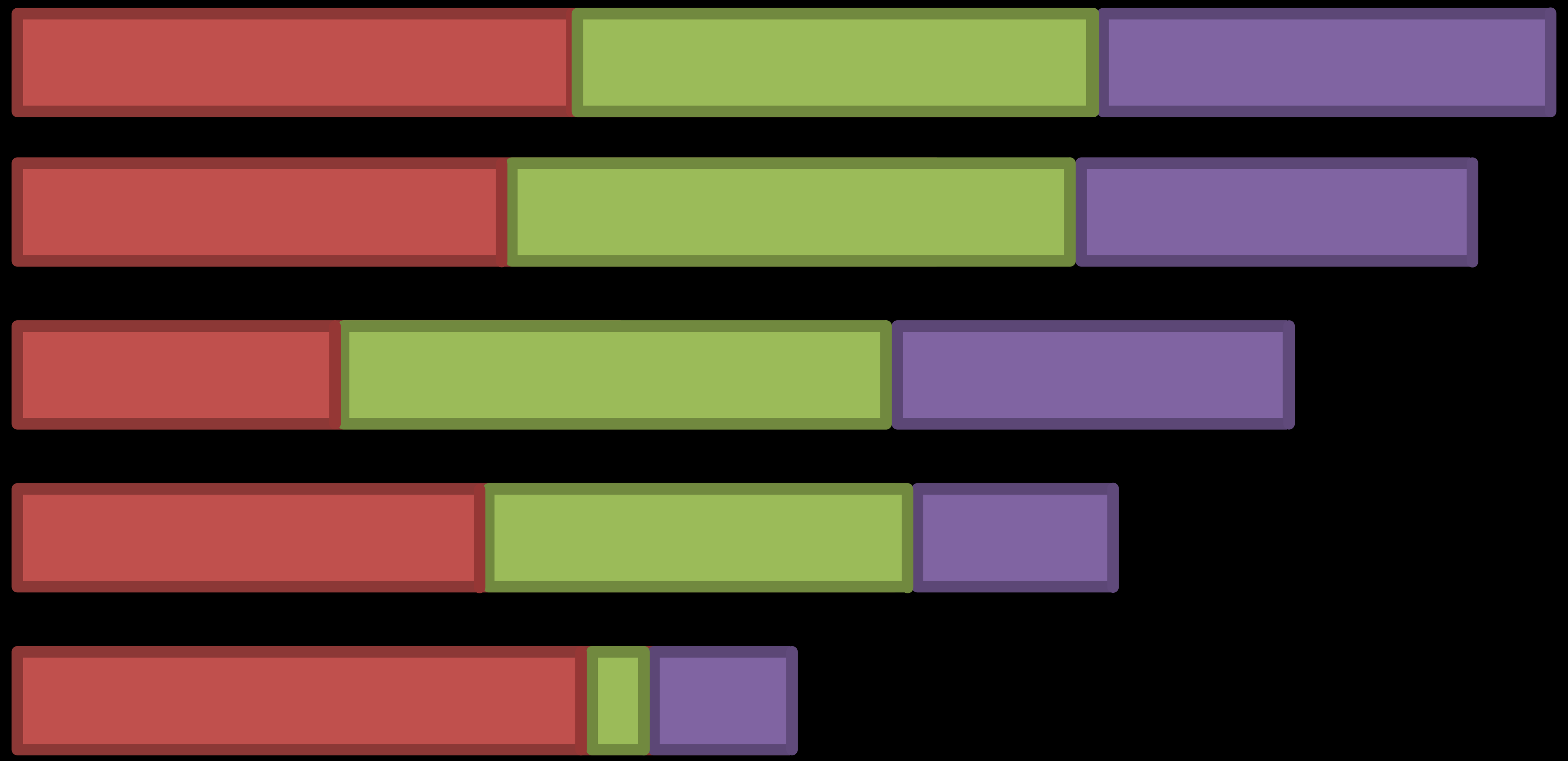
Rank	University	A	B	C
1.	MIT, USA			
2.	Harvard, USA			
3.	Princeton, USA			
4.	Cambridge, UK			
5.	Oxford, UK			



Serial Combiner (as Stacked Bar)

$w_a A + w_b B + w_c C$

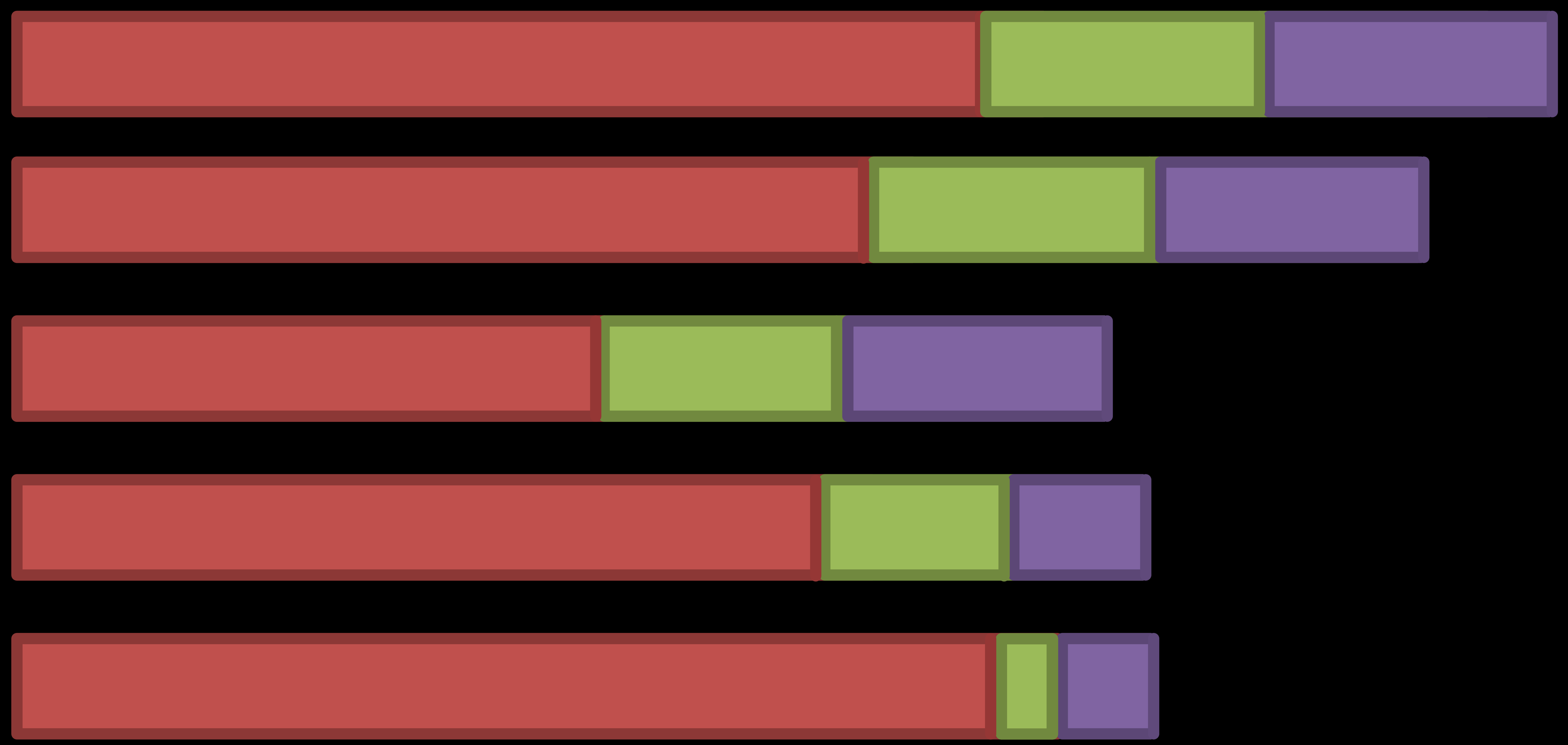
Rank	University	A	B	C
1.	MIT, USA			
2.	Harvard, USA			
3.	Princeton, USA			
4.	Cambridge, UK			
5.	Oxford, UK			



Serial Combiner (as Stacked Bar)

$w_a A + w_b B + w_c C$

Rank	University	A	B	C
1.	MIT, USA			
2.	Harvard, USA			
3.	Princeton, USA			
4.	Cambridge, UK			
5.	Oxford, UK			



10 Requirements

1. Encode Rank
2. Encode Cause of Rank
3. Support Multiple Attributes

Interactive Refinement and Visual Feedback

10 Requirements

1. Encode Rank
2. Encode Cause of Rank
3. Support Multiple Attributes
4. Interactive Refinement and Visual Feedback

Interactive Refinement
and Visual Feedback

128 visible of 906 (14.13%)

Ran	School Name	Country	Faculty/student ratio	Employer reputation	Citations per faculty
	Filter: <None>	Filter: 2 out of 72			
1.	American University	United States			
2.	Arizona State University	United States			
3.	Aston University	United Kingdom			
4.	Birkbeck College, University of L	United Kingdom			
5.	Boston College	United States			
6.	Boston University	United States			
7.	Brandeis University	United States			
8.	Brown University	United States			
9.	Brunel University	United Kingdom			
10.	California Institute of Technology	United States			
11.	Cardiff University	United Kingdom			
12.	Case Western Reserve University	United States			
13.	City University London	United Kingdom			
14.	College of William & Mary	United States			
15.	Colorado State University	United States			
16.	Columbia University	United States			
17.	Cornell University	United States			
18.	Cranfield University	United Kingdom			
19.	Dartmouth College	United States			
20.	Drexel University	United States			
21.	Duke University	United States			
22.	Durham University	United Kingdom			

10 Requirements

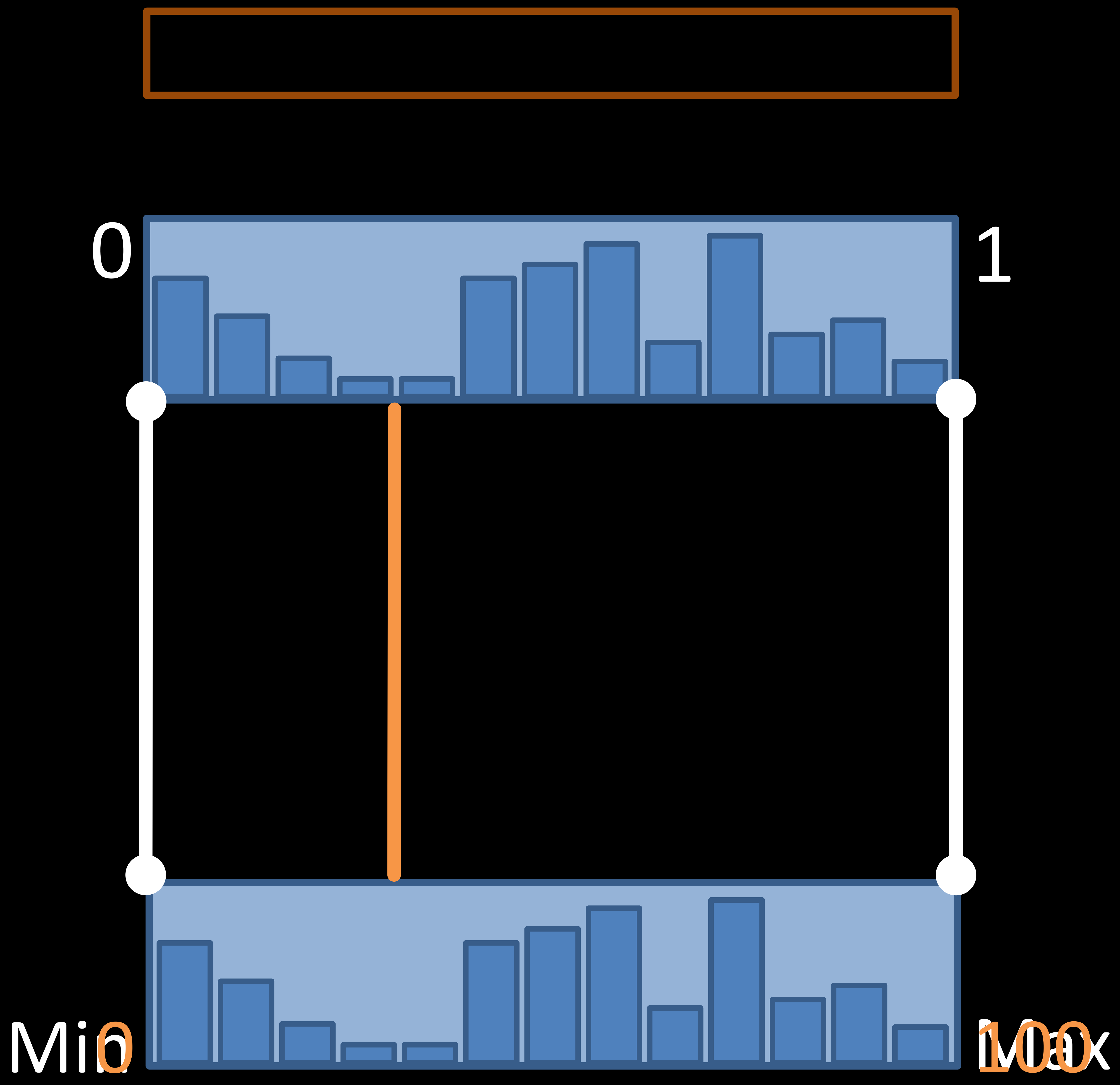
1. Encode Rank
2. Encode Cause of Rank
3. Support Multiple Attributes
4. Interactive Refinement and Visual Feedback

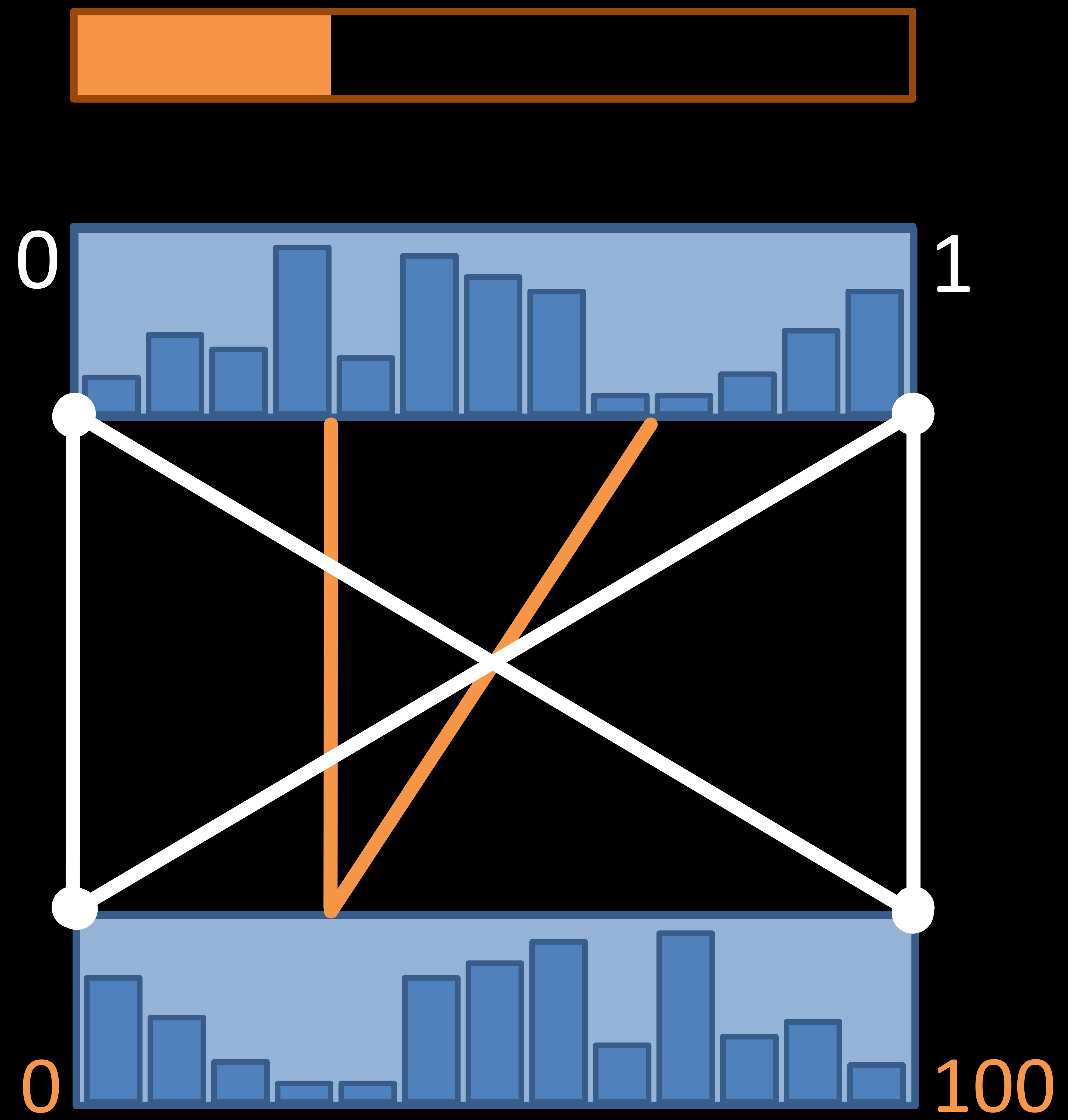
Flexible Mapping of Attributes to Scores

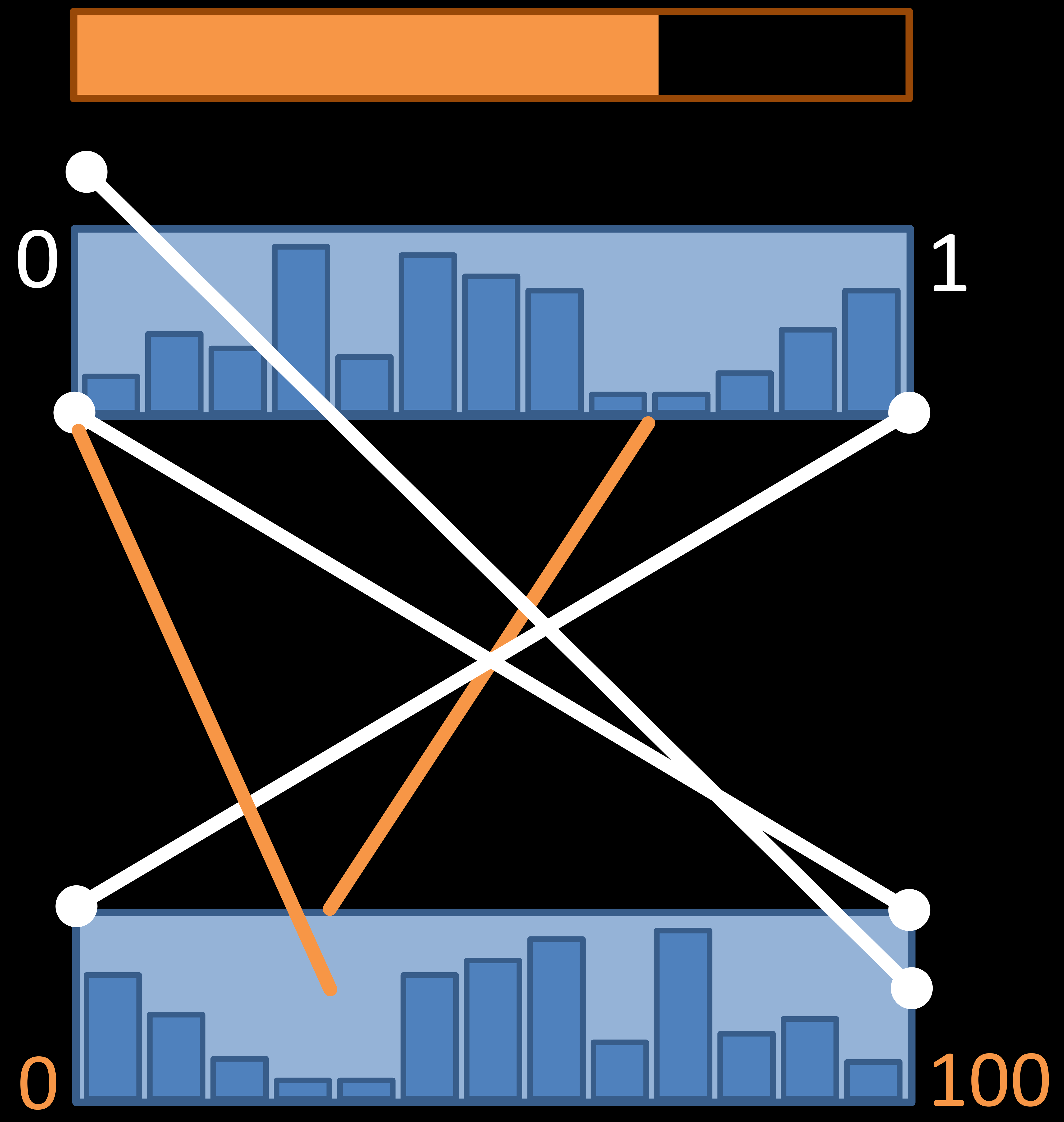
Flexible Mapping of Attributes to Scores

10 Requirements

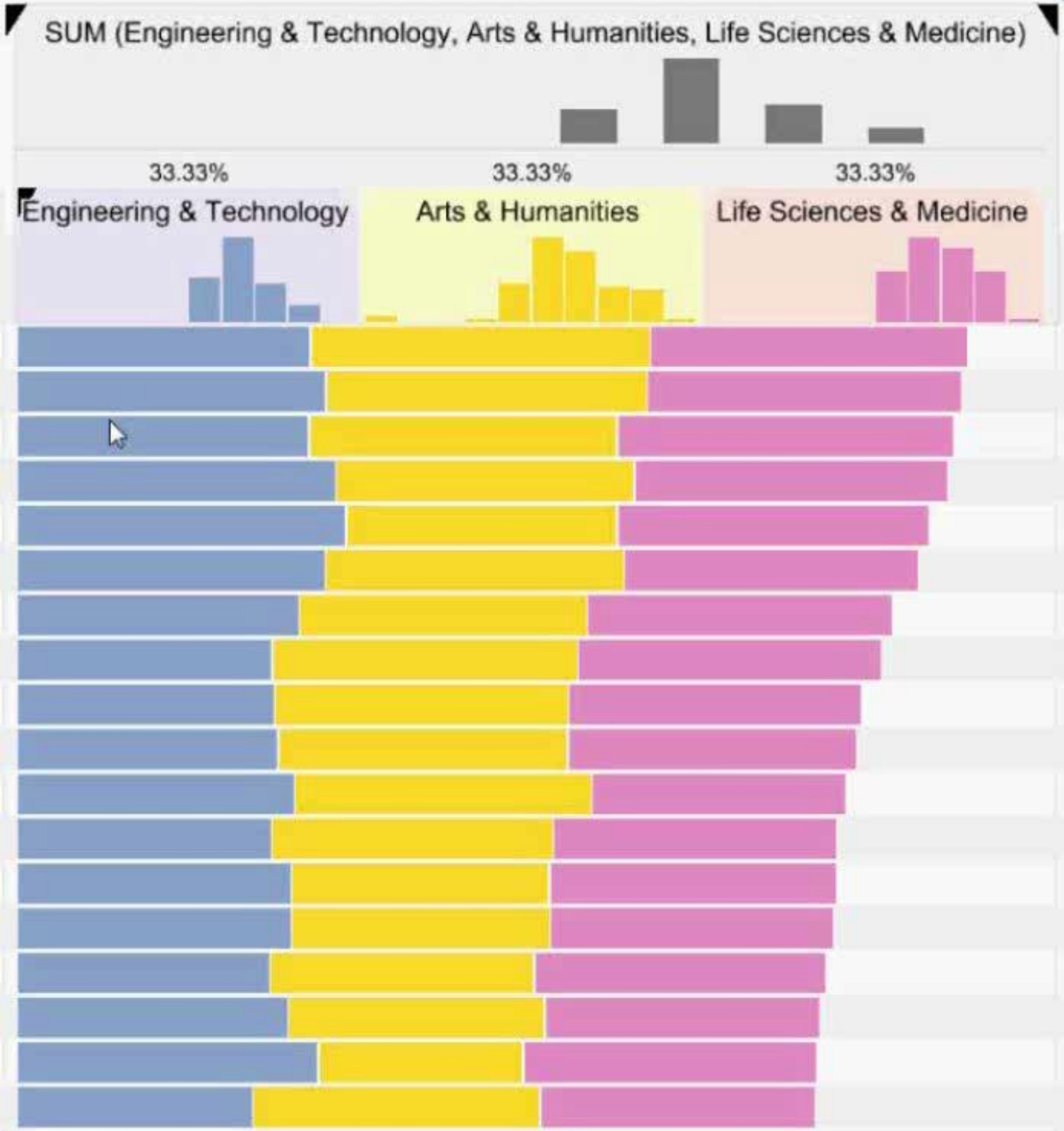
1. Encode Rank
2. Encode Cause of Rank
3. Support Multiple Attributes
4. Interactive Refinement and Visual Feedback
5. Flexible Mapping of Attributes to Scores


0
1
Min
0
Max
100


0
1
0
100


0
1
0
100

93 visible of 399 (23.31%)



Ran	School Name Filter: <None>	Country Filter: 2 out of 43
1.	University of Oxford	United Kingdom
2.	University of Cambridge	United Kingdom
3.	Harvard University	United States
4.	Stanford University	United States
5.	Massachusetts Institute of Technology (MIT	United States
6.	University of California, Berkeley (UCB)	United States
7.	University of California, Los Angeles (UCL	United States
8.	Yale University	United States
9.	UCL (University College London)	United Kingdom
10.	Columbia University	United States
11.	Princeton University	United States
12.	University of Edinburgh	United Kingdom
13.	University of Michigan	United States
14.	Cornell University	United States
15.	University of Pennsylvania	United States
16.	The University of Manchester	United Kingdom
17.	Imperial College London	United Kingdom
18.	University of Chicago	United States

10 Requirements

1. Encode Rank
2. Encode Cause of Rank
3. Support Multiple Attributes
4. Interactive Refinement and Visual Feedback
5. Flexible Mapping of Attributes to Scores

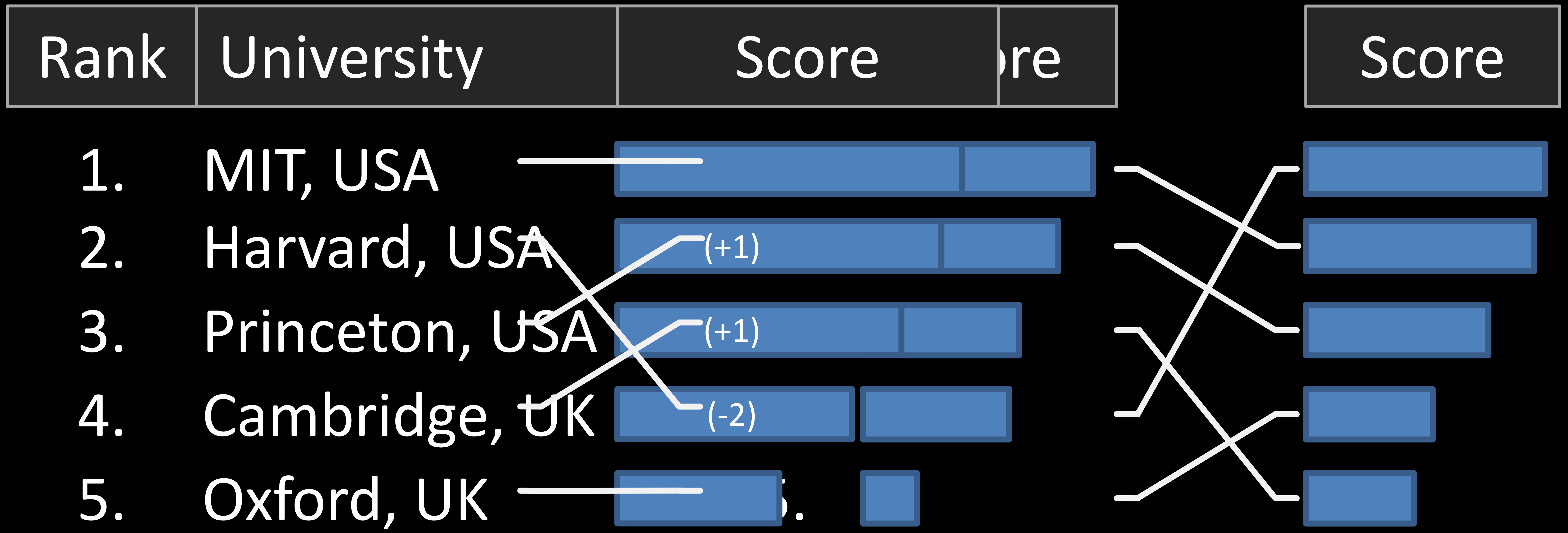
Compare Rankings

10 Requirements

1. Encode Rank
2. Encode Cause of Rank
3. Support Multiple Attributes
4. Interactive Refinement and Visual Feedback
5. Flexible Mapping of Attributes to Scores
6. Compare Rankings

Compare Rankings

Bump Charts


Rank
University
Score
Score
1. MIT, USA
2. Harvard, USA
3. Princeton, USA
4. Cambridge, UK
5. Oxford, UK
(+1)
(+1)
(-2)

Bump Charts


Rank
University
Score
1.
MIT, USA
2.
Harvard, USA
3.
Princeton, USA
4.
Cambridge, UK
5.
Oxford, UK
Rank
Score
1.
(+1)2.
(+1)3.
(-2)4.
5.
Score


453 visible of 453 (100.00%)
World University Ranking
46.55%
32.72%
20.72%
Rank
School Name
MAX (Academic re
Faculty/stud
Citatio
Filter:
<None>
1. Massachusetts Institute of Te
2. California Institute of Technol
3. Harvard University
4. University of Cambridge
5. UCL (University College Lond
6. University of Oxford
7. Princeton University
8. Imperial College London
9. University of Chicago
10. Stanford University
11. Columbia University
12. Duke University
13. University of Pennsylvania
14. Johns Hopkins University
15. Yale University
16. University of Michigan
17. Ecole normale supérieure, Pa
18. Northwestern University

10 Requirements

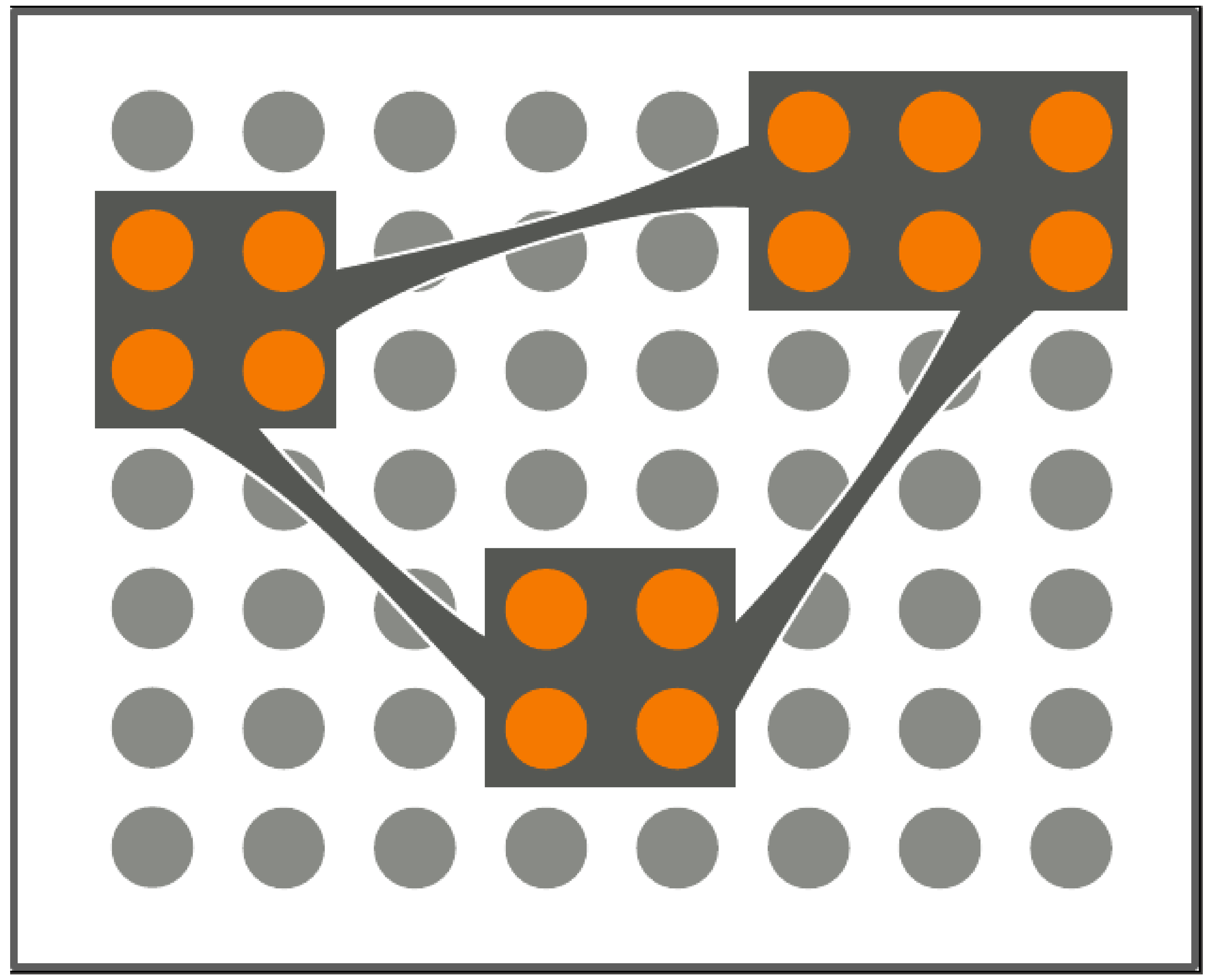
1. Encode Rank
2. Encode Cause of Rank
3. Support Multiple Attributes
4. Interactive Refinement and Visual Feedback
5. Flexible Mapping of Attributes to Scores
6. Compare Rankings
7. Scalability
8. Filtering
9. Handle Missing Values
10. Optimization

Demos, Videos & More:

http://lineup.caleydo.org

WHAT IS NEXT?

Large & Heterogeneous Data



Research Approach: Semantic Subsets

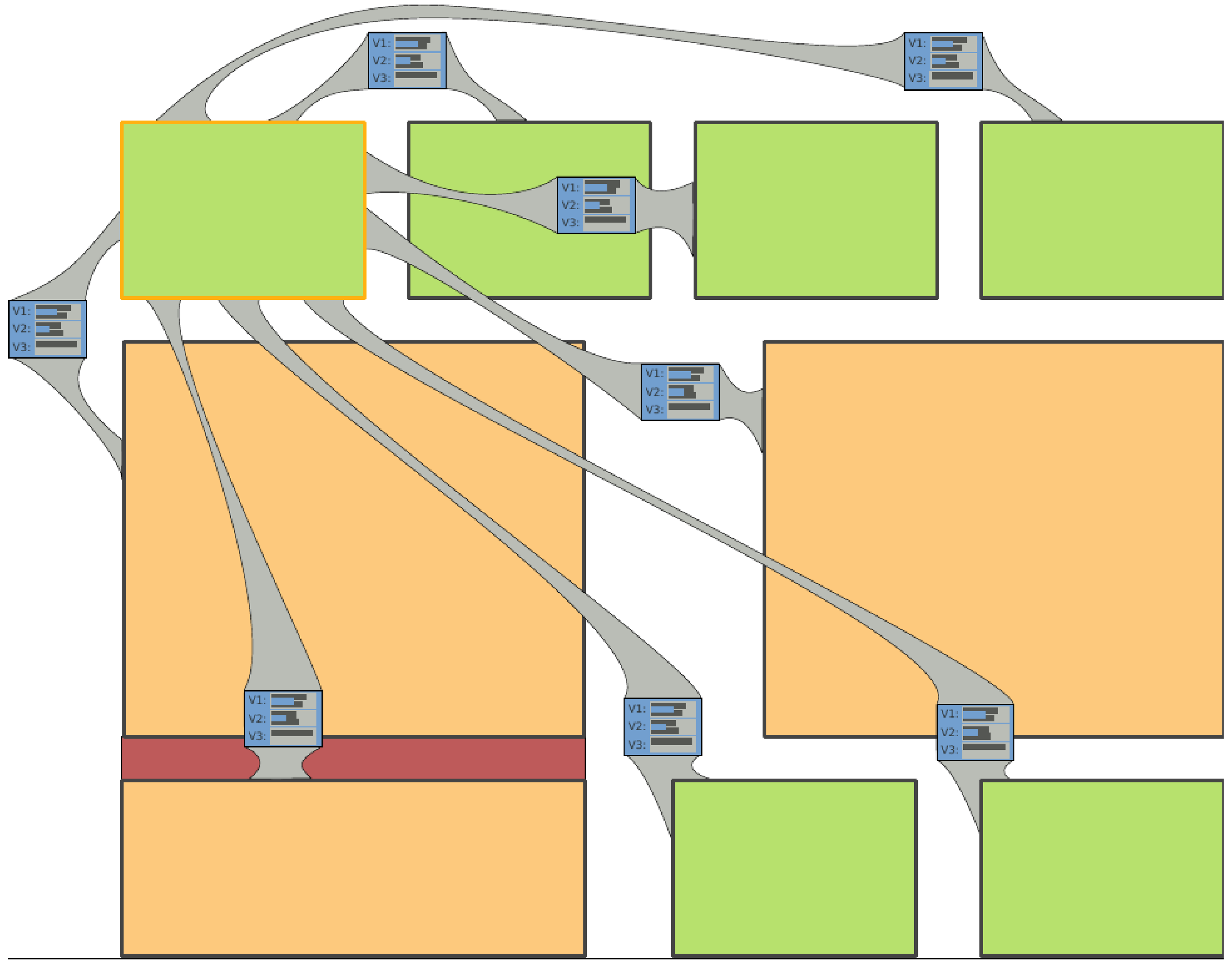
Interpretable Division (why is this a subset?)

Representing Data on Multiple Levels of Detail

Visualizing Relationships

Ensuring Coverage

Uncertainty and Ambiguity



High-Impact Domain Problems

?

Visualization Approaches for Biomolecular Data

Alexander Lex, Harvard University
alex@seas.harvard.edu
http://alexander-lex.com
@alexander_lex

Credits:
Marc Streit, Nils Gehlenborg, Hanspeter Pfister, Anne Mai Wasserman, Mark Borowsky, Christian Partl, Denis Kalkofen, Samuel Gratzl, Dieter Schmalstieg

HARVARD
School of Engineering and Applied Sciences