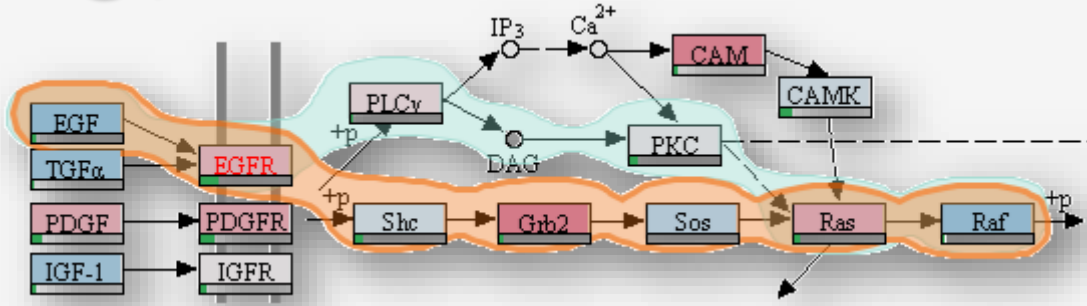
Data Visualization in Molecular Biology

Alexander Lex



July 29, 2013

CALEYDO

HARVARD
School of Engineering and Applied Sciences

research requires
understanding data

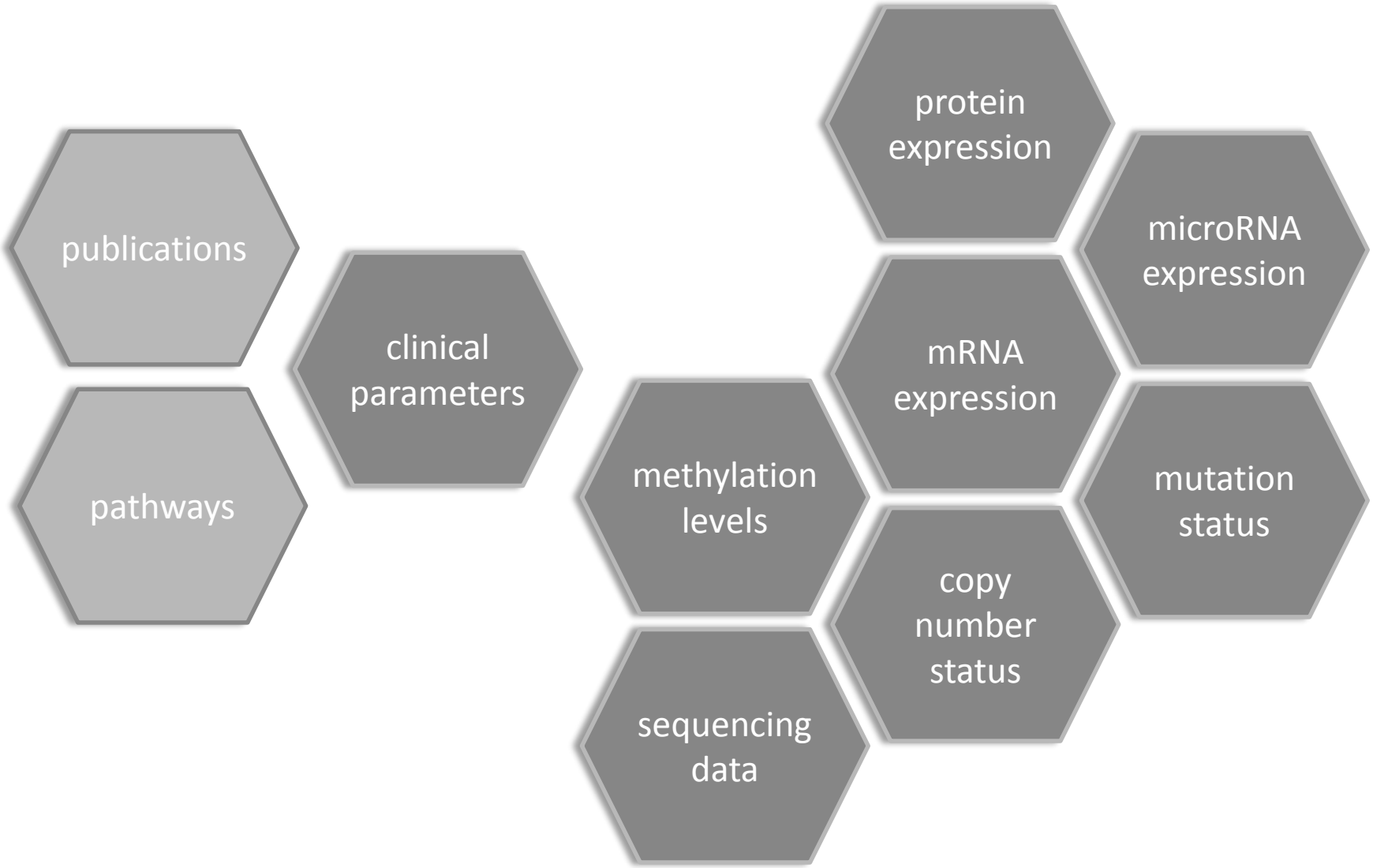
but there is so much of it…

What is important?

Where are the connections?

The Cancer Genome Atlas

Understanding genomics to improve cancer care



Data Visualization

… makes the data accessible

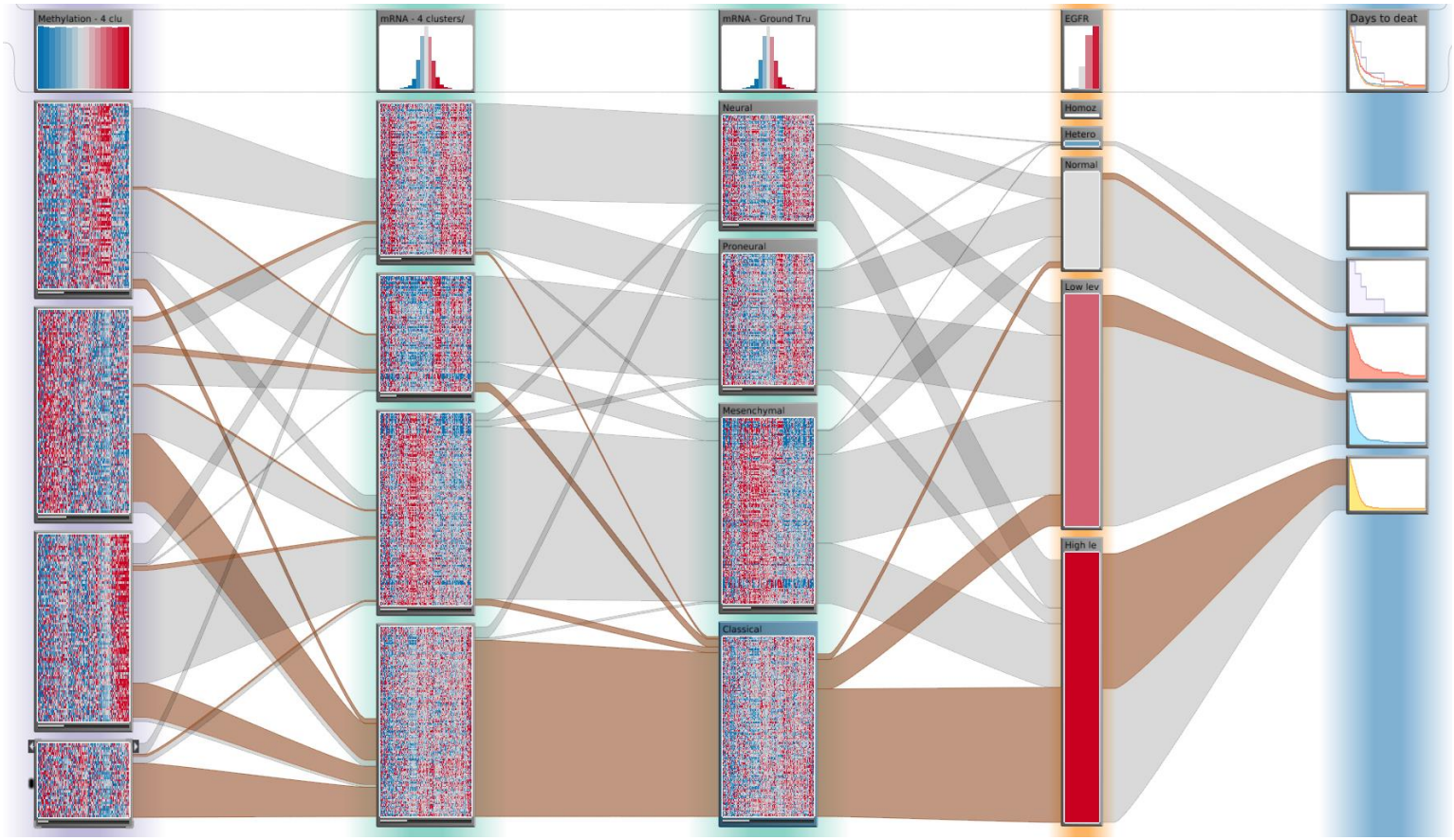
… combines strengths of humans and computers

… enables insight

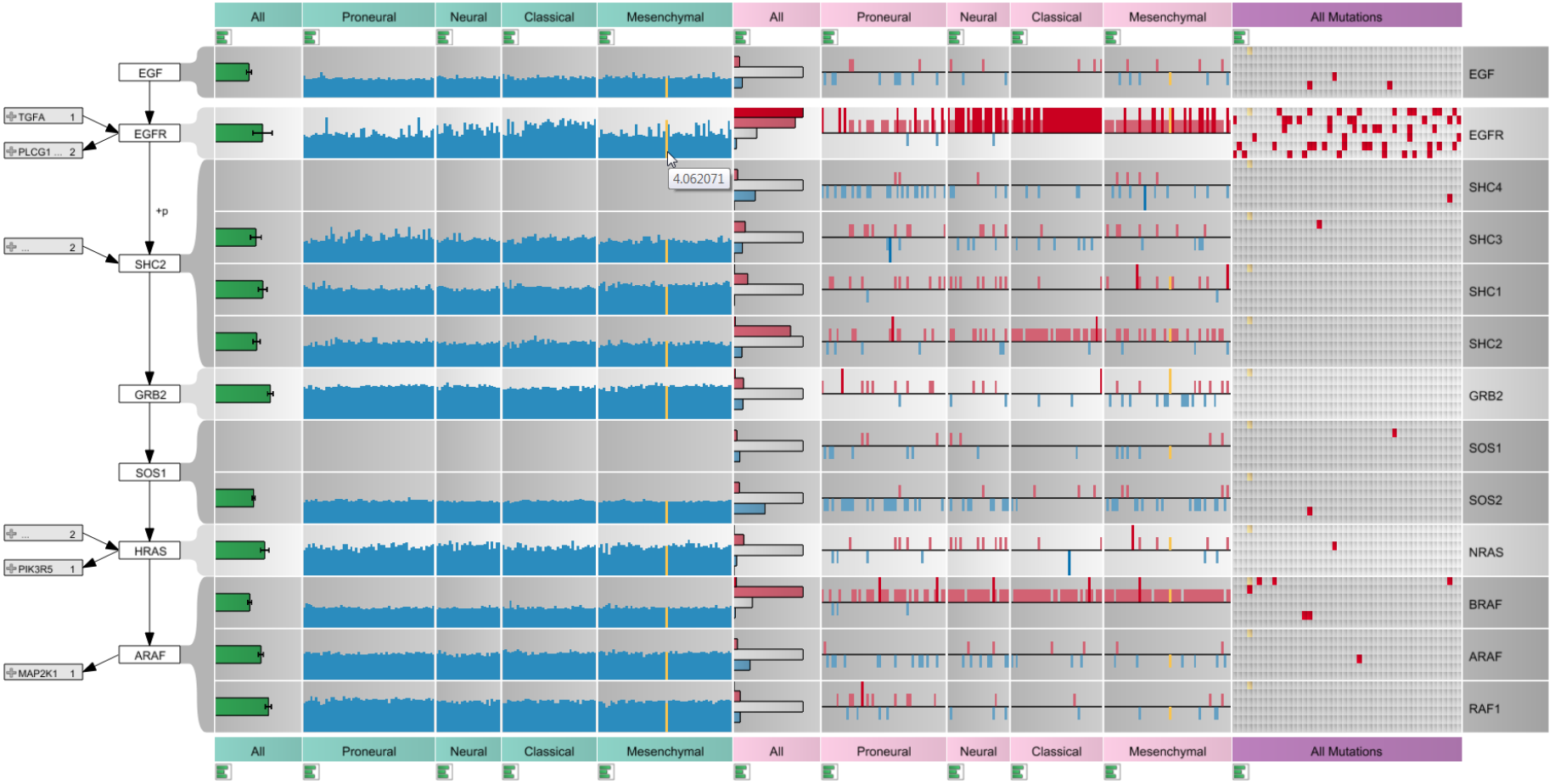
… communicates

THREE TOPICS

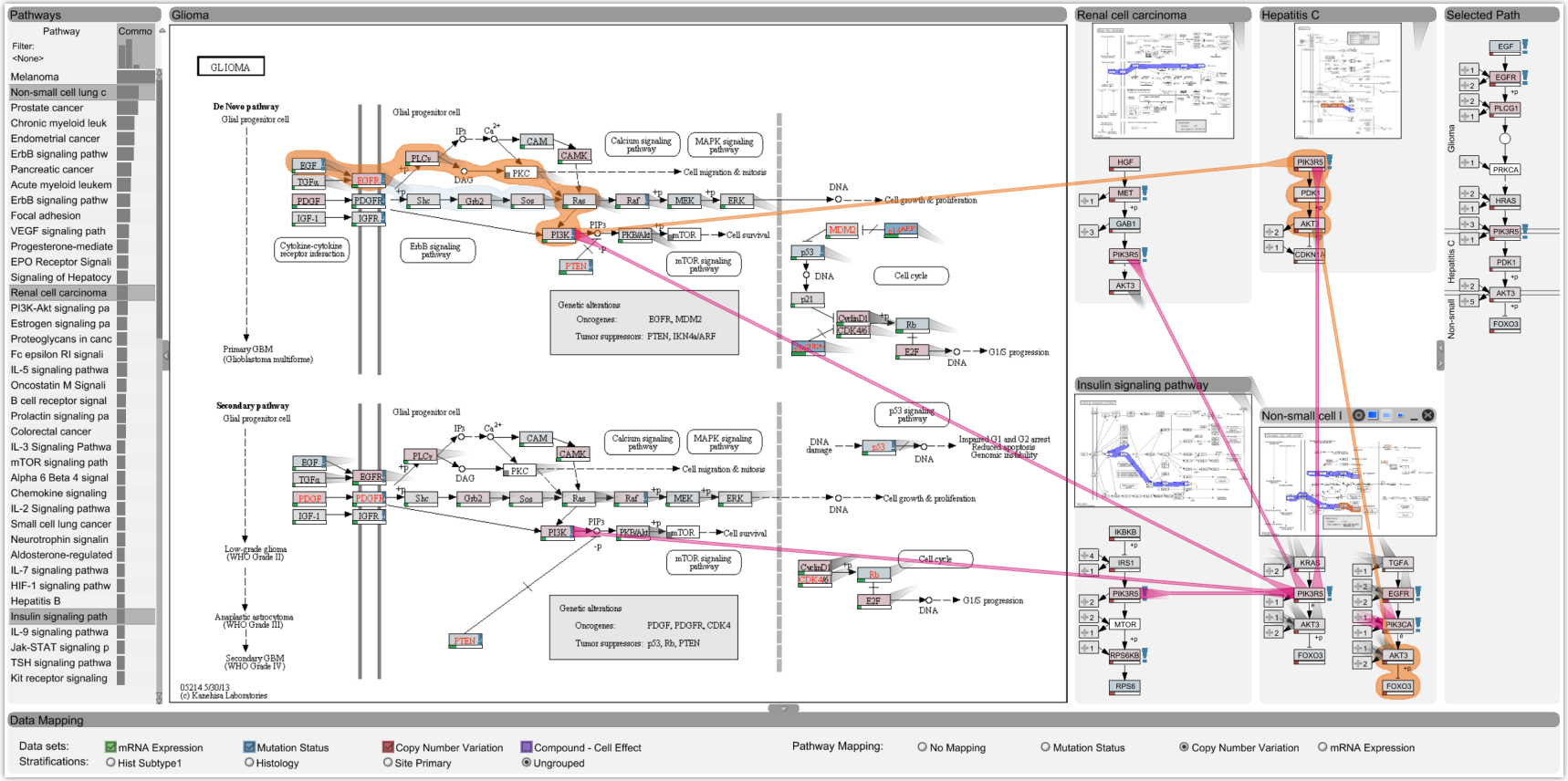
Cancer Subtype Analysis



Pathways & Experimental Data



Managing Pathways & Cross-Pathway Analysis

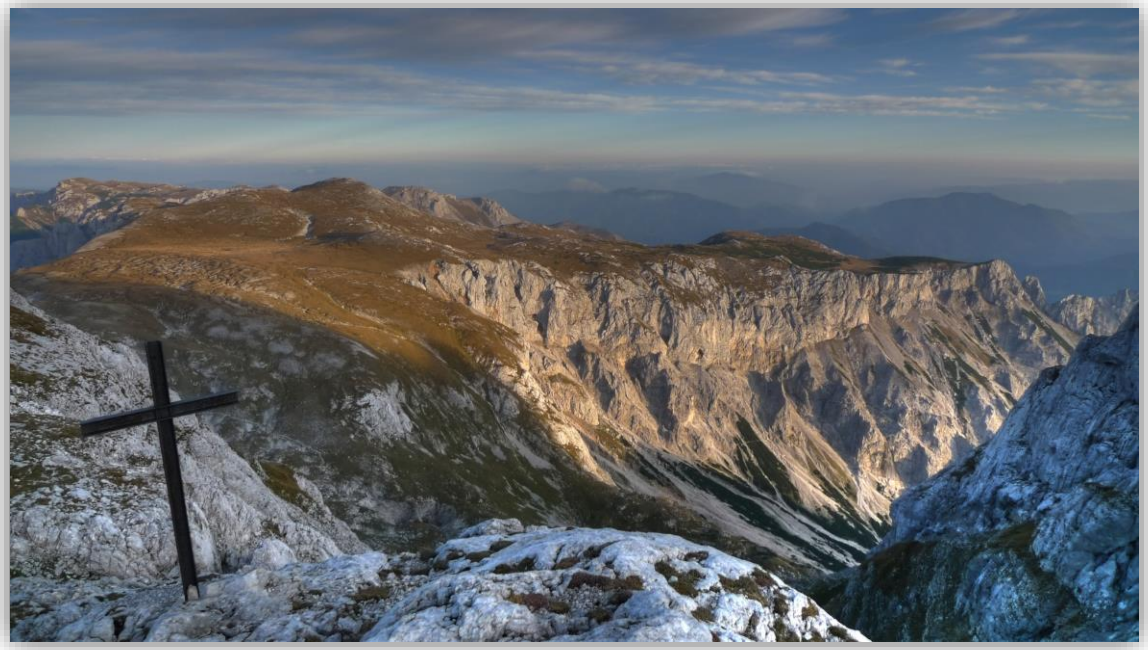


Who am I?

PostDoc @ Harvard,
Hanspeter Pfister's Group

PhD from TU Graz, Austria

Co-leader of
Caleydo Project



What is Caleydo?

Software analyzing molecular biology data

Software for doing research in visualization

- developed in academic setting
- platform for trying out radically new visualization ideas

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

CALEYDO

What is Caleydo?

Open source platform for developing visualization and data analysis techniques

- easily extendible due to plug-in architecture
- you can create your own views
- you can plug-in your own algorithms

CALEYDO

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*

TU Graz

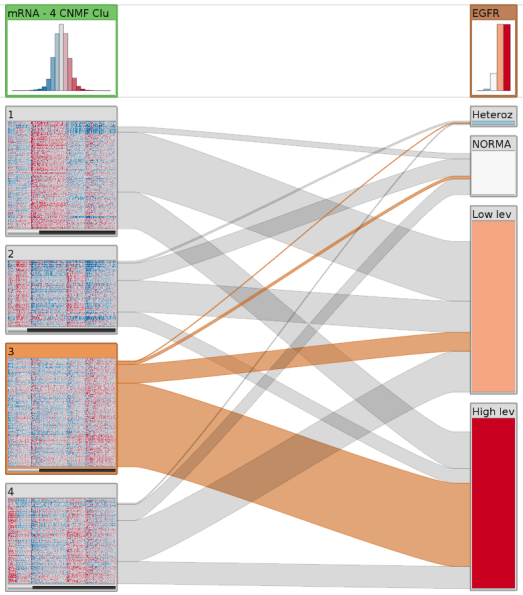
JKU JOHANNES KEPLER UNIVERSITÄT LINZ

HARVARD School of Engineering and Applied Sciences

HARVARD MEDICAL SCHOOL

Caleydo StratomeX

CANCER SUBTYPE VISUALIZATION



Cancer Subtypes

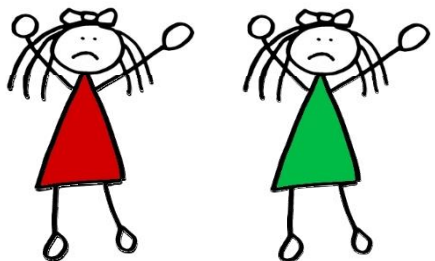
Cancer types are not homogeneous

They are divided into **Subtypes**

- different histology
- different molecular alterations

Subtypes have serious implications

- different treatment for subtypes
- prognosis varies between subtypes



Cancer Subtype Analysis

Done using *many different types of data,*

for *large numbers of patients.*

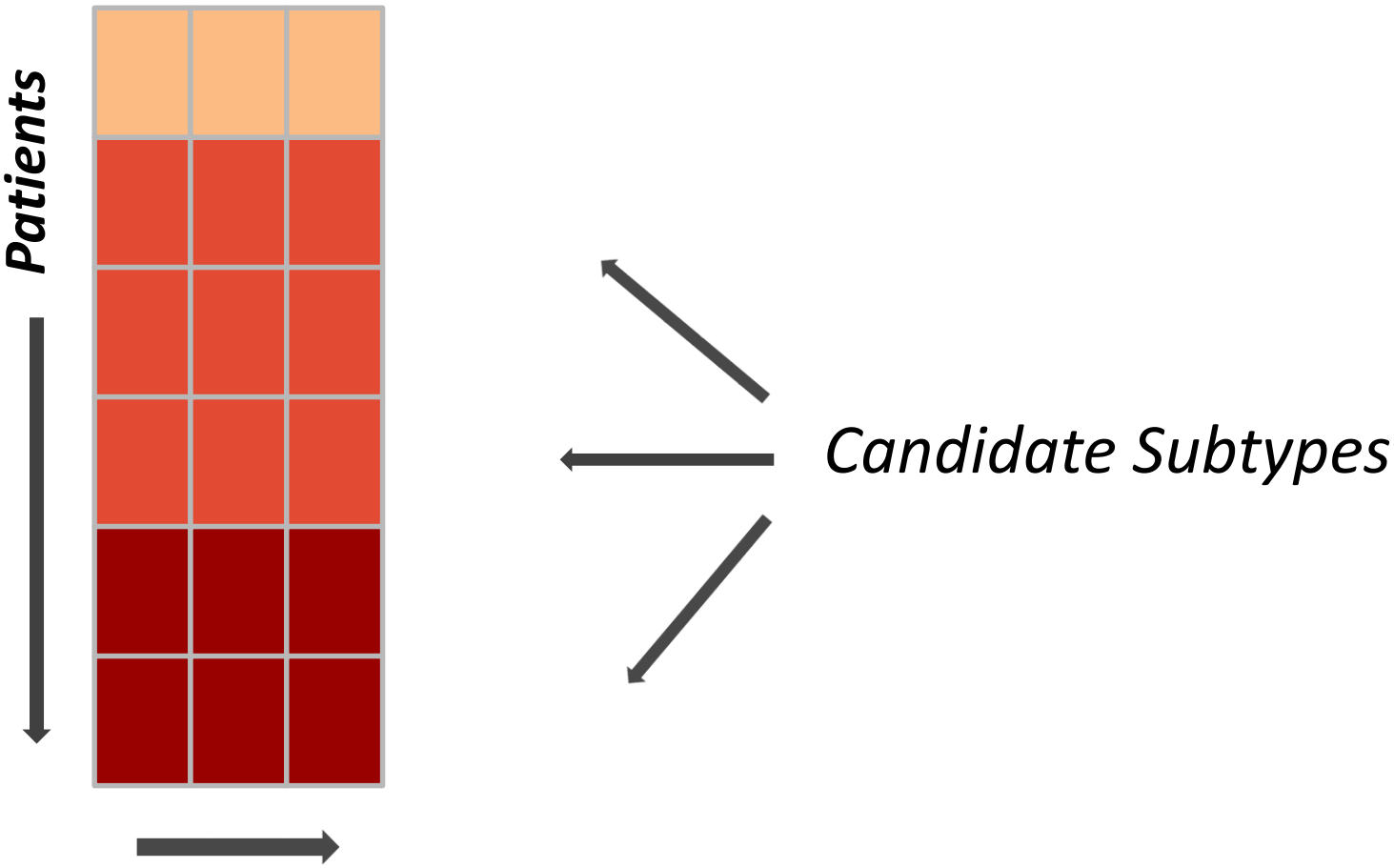
Goal:

Support tumor subtype characterization
through

Integrative visual analysis
of cancer genomics data sets.

Tabular Data

Stratification



Stratification of a Single Dataset

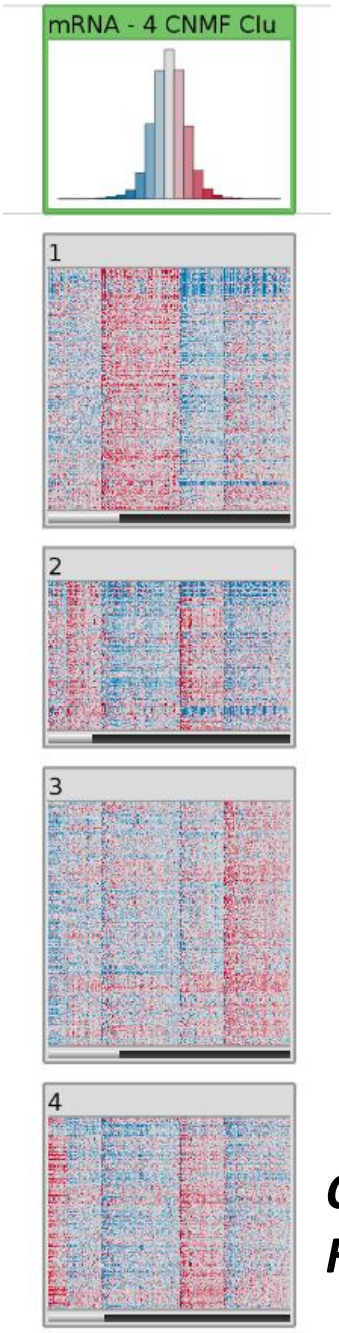
Subtypes are identified
by stratifying datasets, e.g.,

based on an expression pattern

a mutation status

a copy number alteration

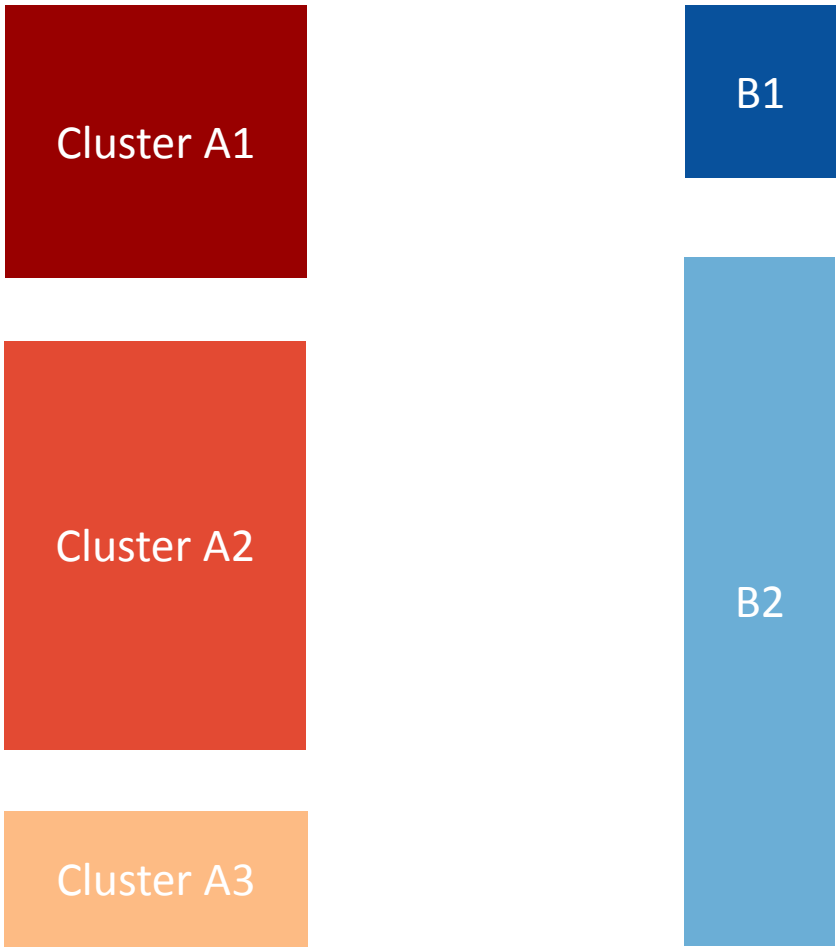
a combination of these



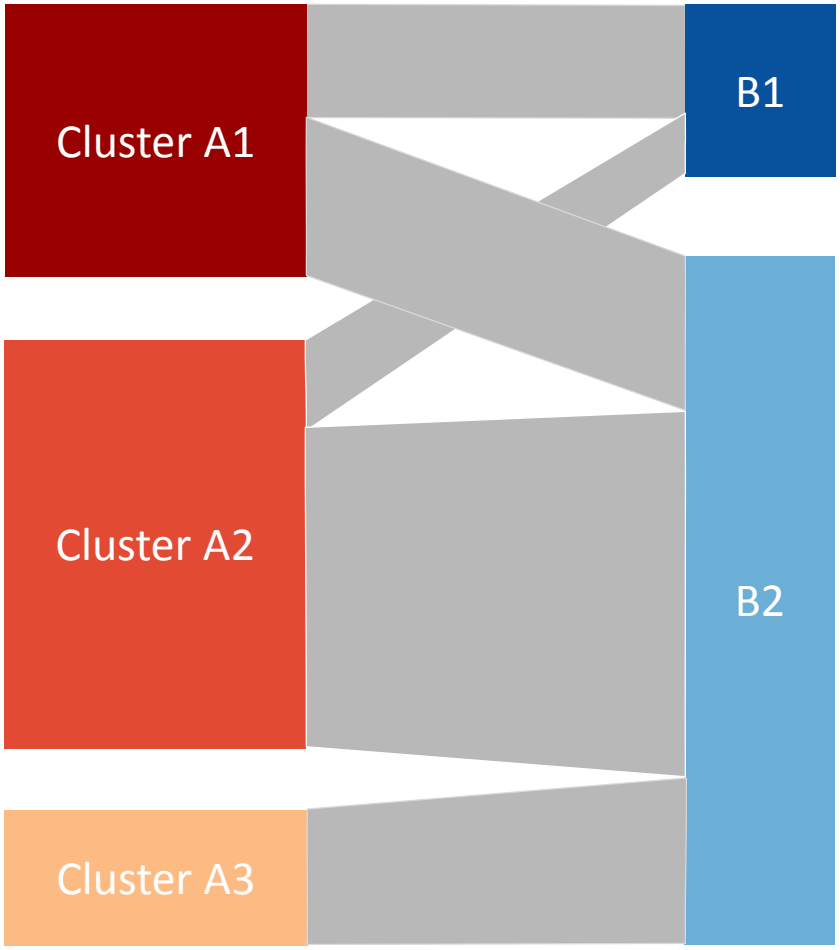
Header / Summary of whole Stratification

Candidate Subtype / Heat Map

Stratification of Multiple Datasets

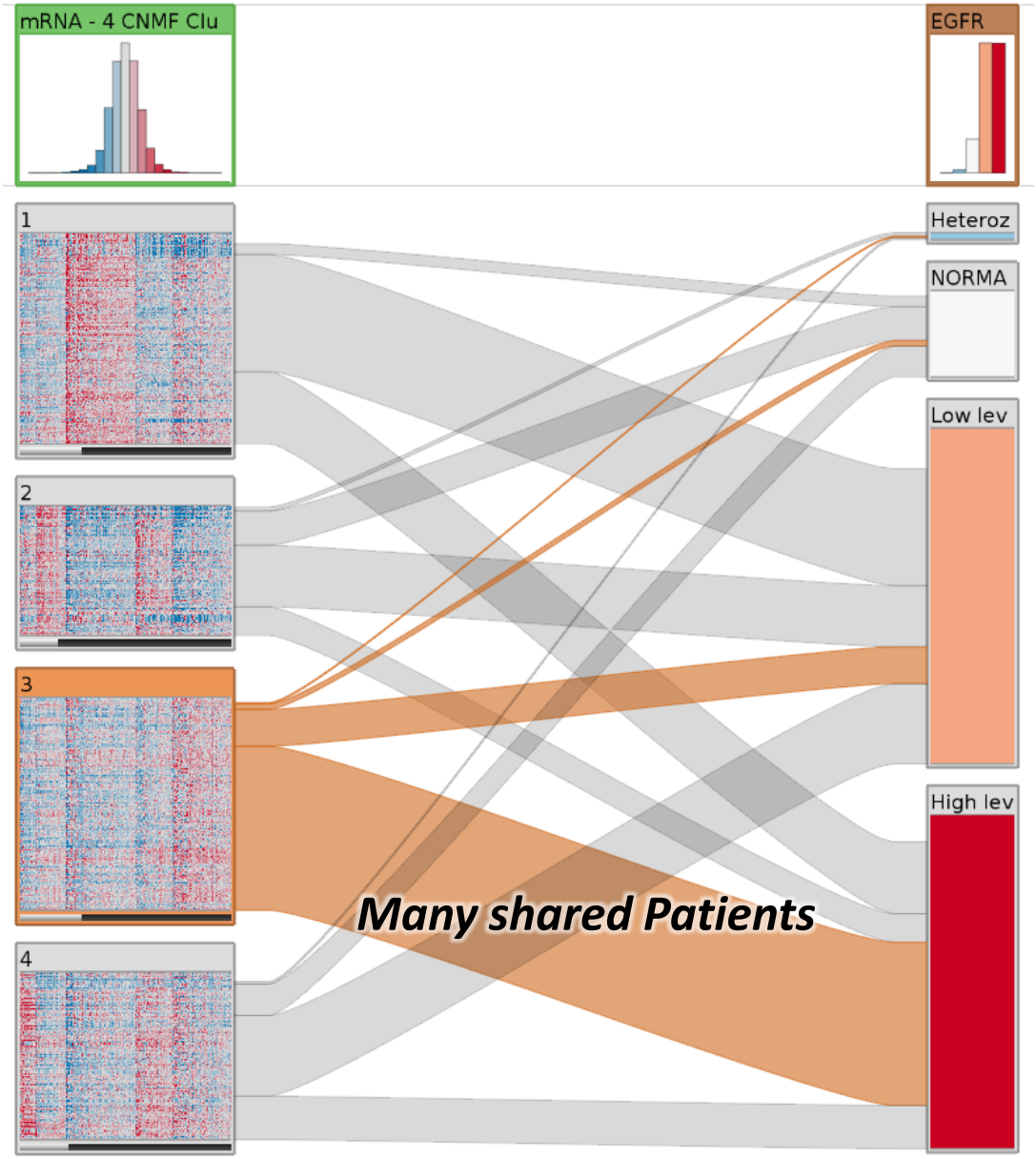


Stratification of Multiple Datasets



Tabular
e.g., mRNA

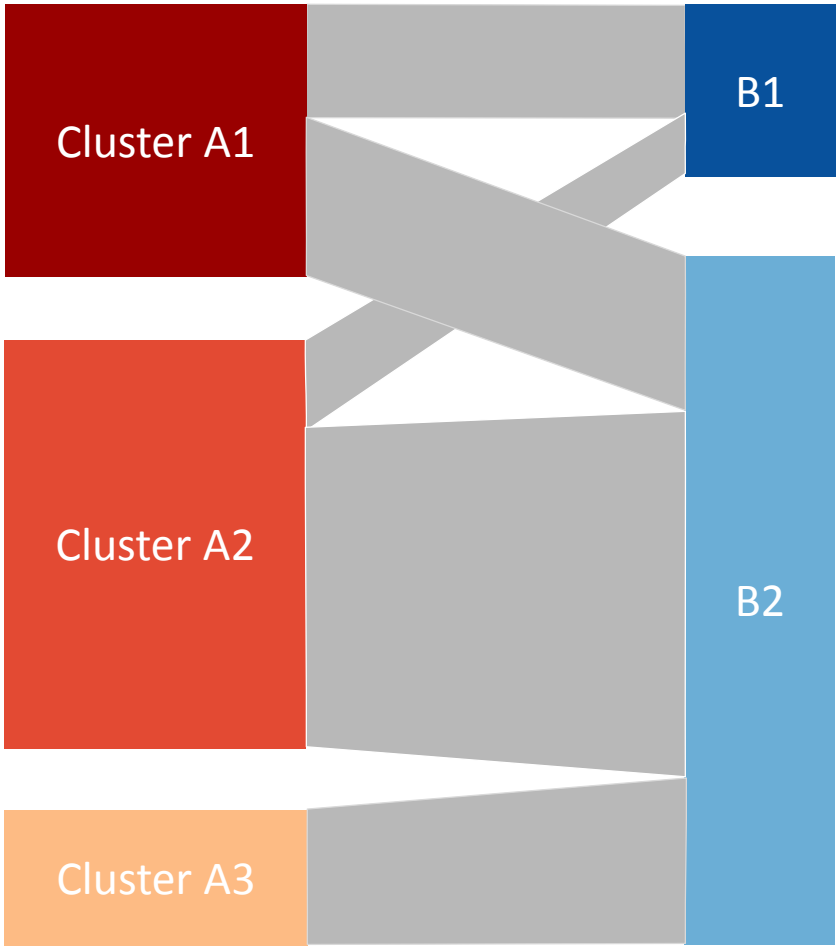
Categorical,
e.g., mutation status



Clustering of mRNA Data

Stratification on Copy Number Status

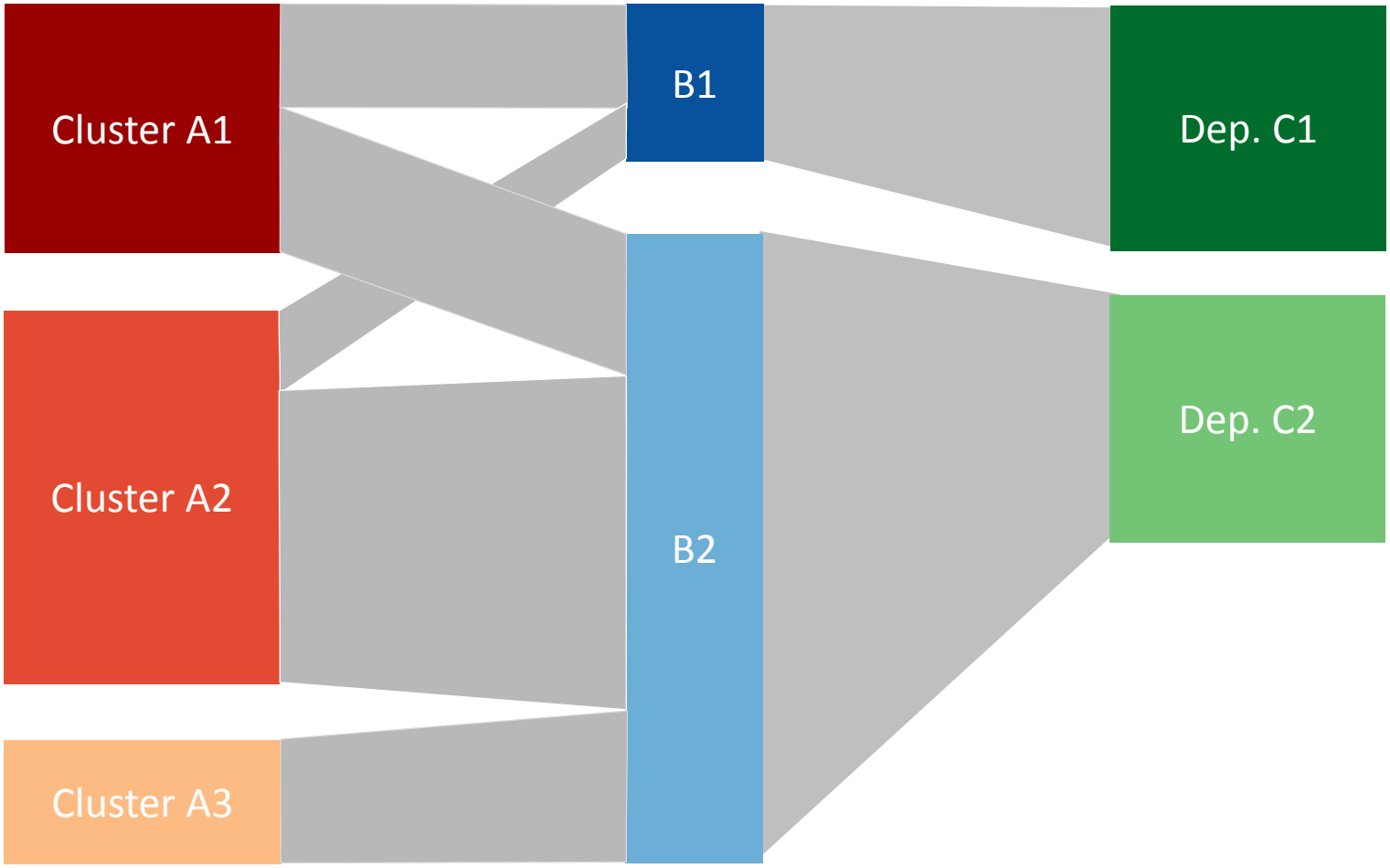
Stratification of Multiple Datasets



Tabular
e.g., mRNA

Categorical,
e.g., mutation status

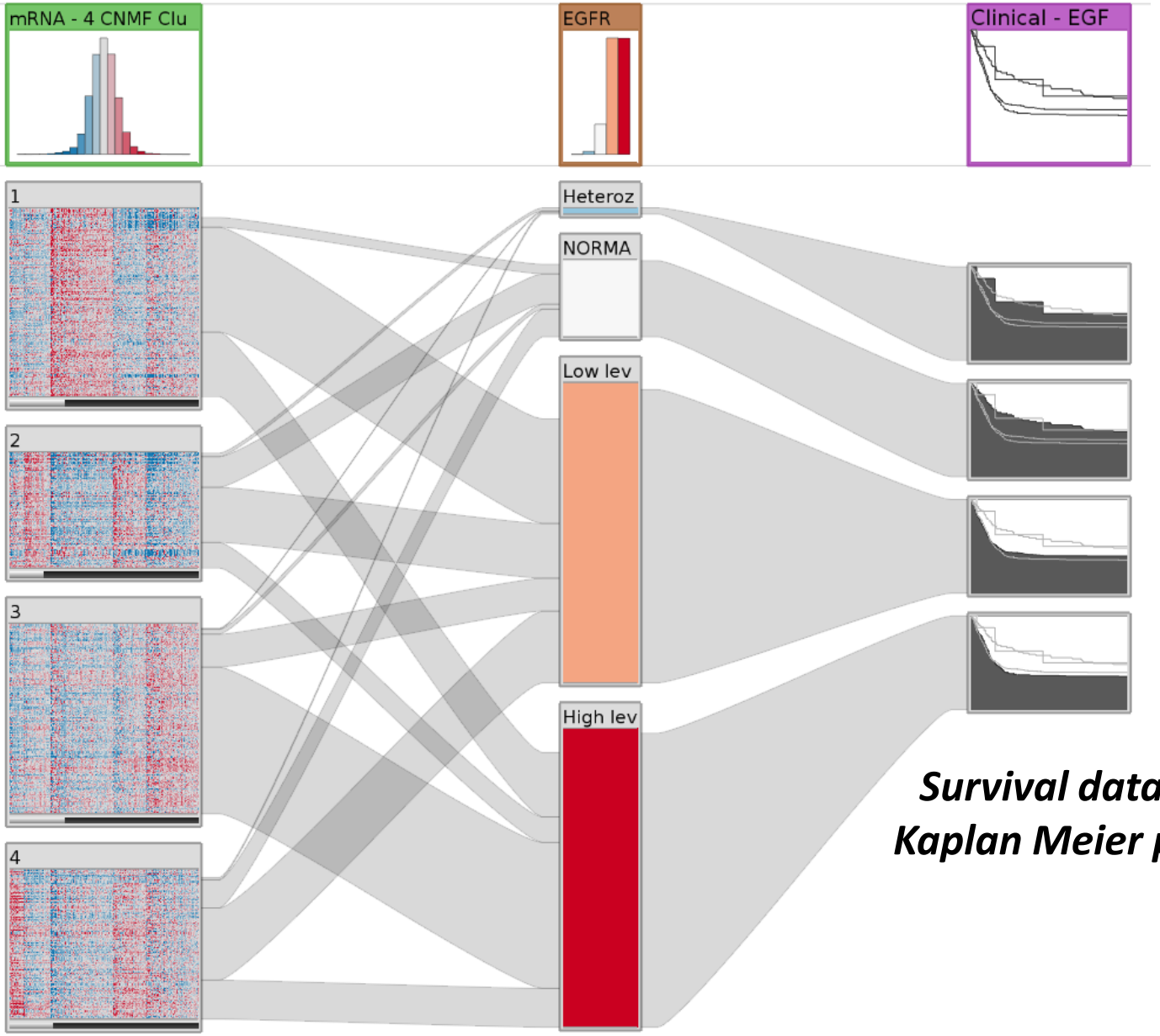
Stratification of Multiple Datasets



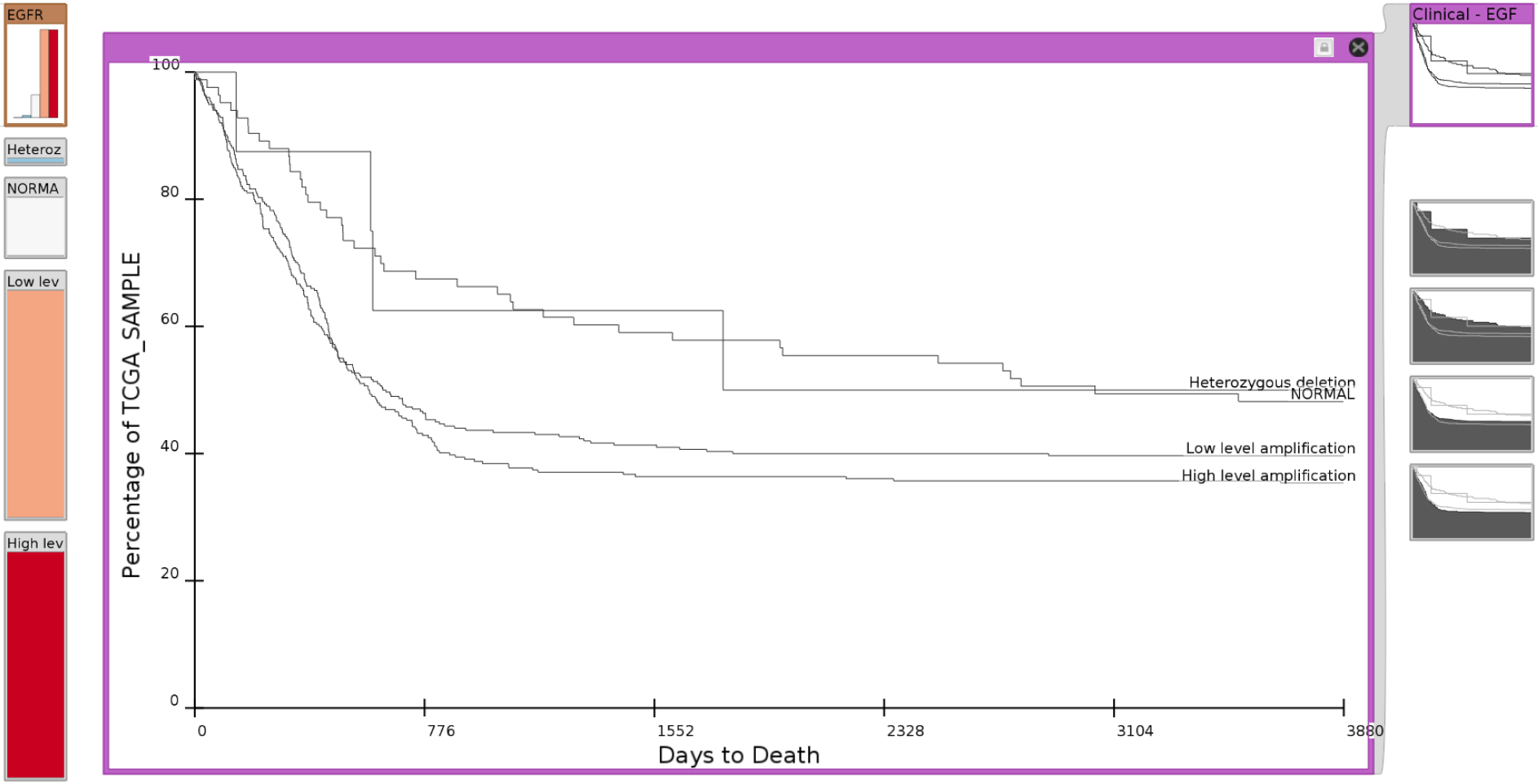
Tabular
e.g., mRNA

Categorical,
e.g., mutation status

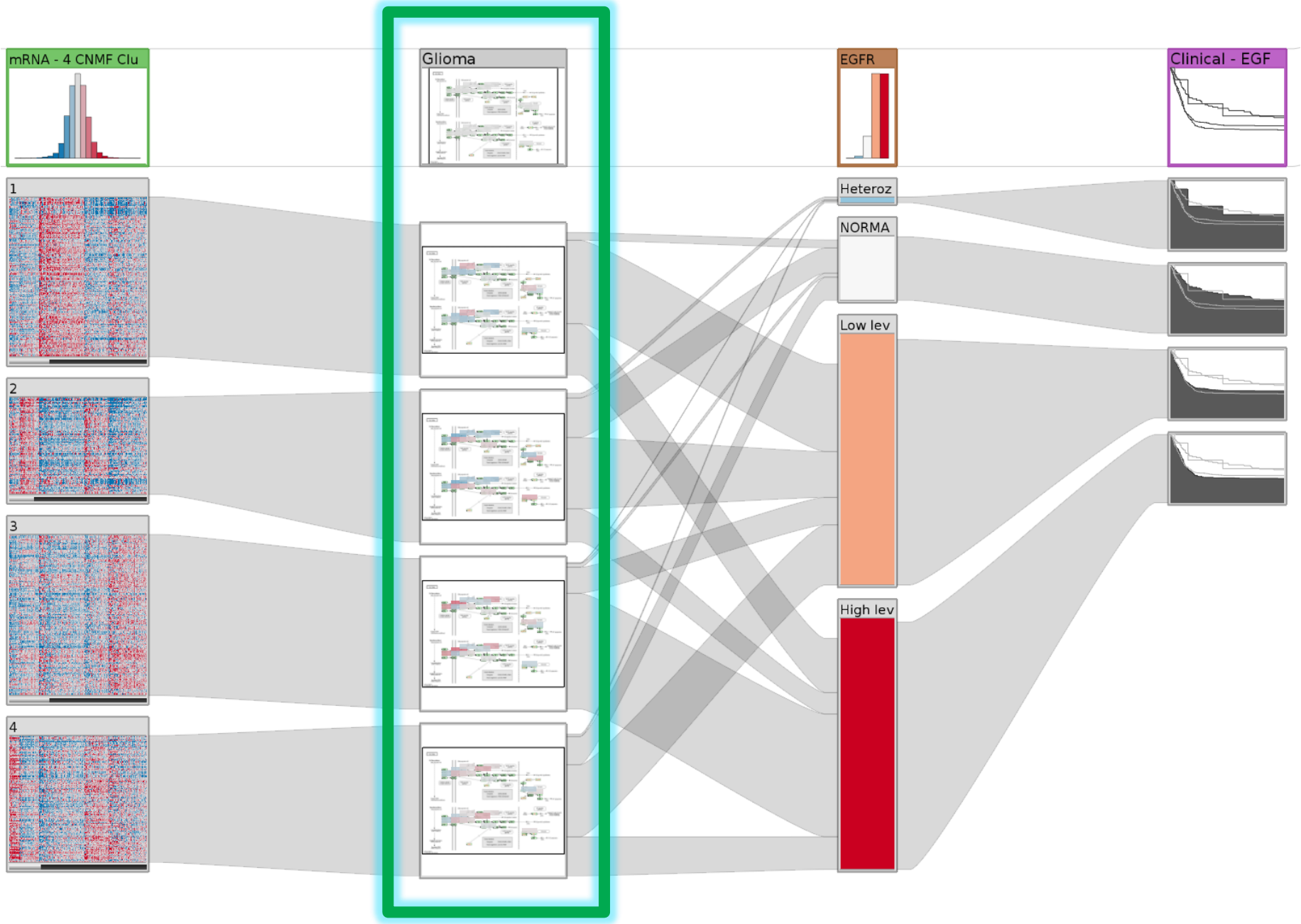
Dependent Data,
e.g. clinical data



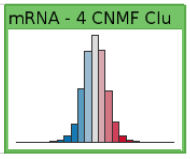
Survival data in Kaplan Meier plots

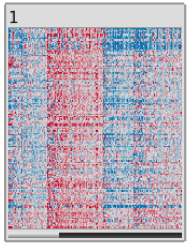

EGFR
Heteroz
NORMA
Low lev
High lev
Clinical - EGF
100
80
60
40
20
0
Percentage of TCGA_SAMPLE
0
776
1552
2328
3104
3880
Days to Death
Heterozygous deletion
NORMAL
Low level amplification
High level amplification

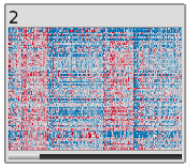
Detail View

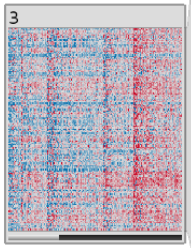


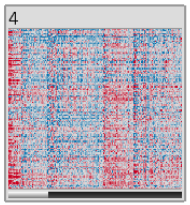
Dependent Pathway

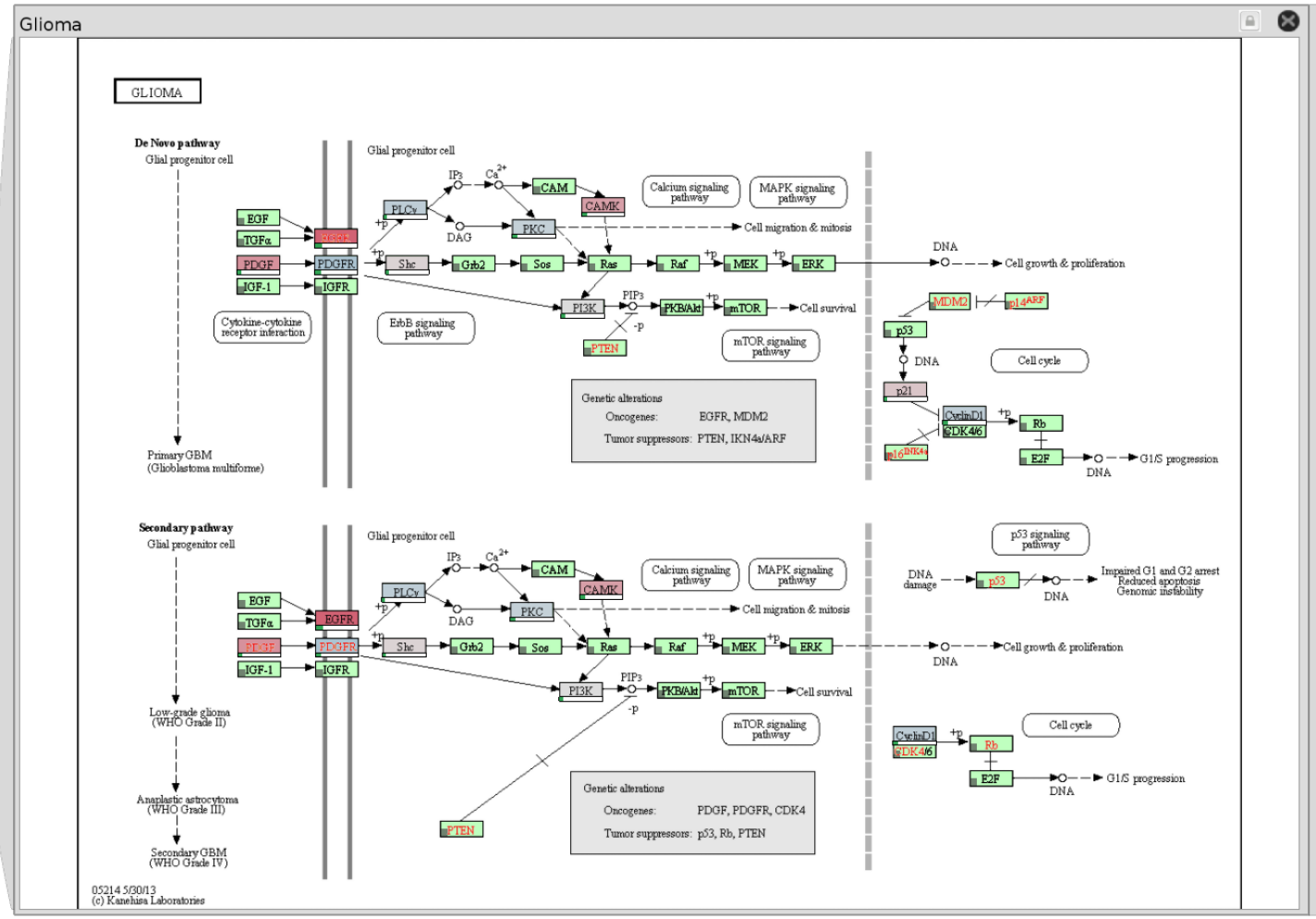

mRNA - 4 CNMF Clu

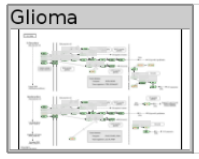

1

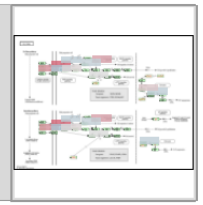

2

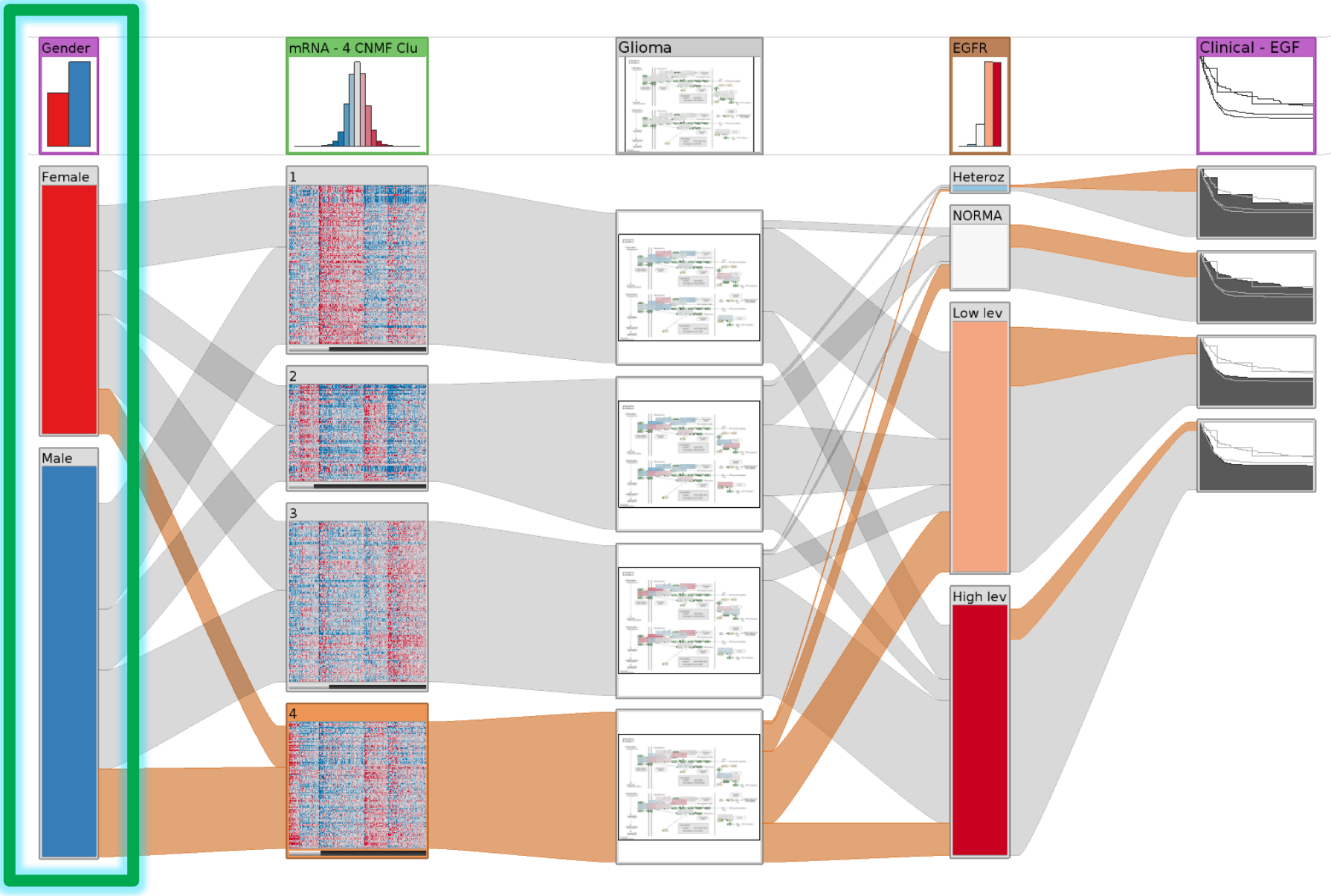

3


4


Glioma
GLIOMA
De Novo pathway
Glial progenitor cell
Primary GBM
(Glioblastoma multiforme)
EGF
TGFα
PDGF
IGF-1
EGFR
PDGFR
IGFR
Cytokine-cytokine receptor interaction
ErbB signaling pathway
PLCγ
IP3
Ca2+
CAM
CAMK
Calcium signaling pathway
MAPK signaling pathway
PKC
DAG
Cell migration & mitosis
Shc
Grb2
Sos
Ras
Raf
MEK
ERK
DNA
Cell growth & proliferation
PI3K
PIP3
PKB/Akt
mTOR
Cell survival
PTEN
mTOR signaling pathway
Genetic alterations
Oncogenes: EGFR, MDM2
Tumor suppressors: PTEN, IKN4a/ARF
MDM2
p14ARF
p53
Cell cycle
p21
CyclinD1
CDK4/6
p16INK4
Rb
E2F
G1/S progression
Secondary pathway
Low-grade glioma
(WHO Grade II)
Anaplastic astrocytoma
(WHO Grade III)
Secondary GBM
(WHO Grade IV)
p53 signaling pathway
DNA damage
p53
Impaired G1 and G2 arrest
Reduced apoptosis
Genomic instability
Genetic alterations
Oncogenes: PDGF, PDGFR, CDK4
Tumor suppressors: p53, Rb, PTEN
05214 5/30/13
(c) Kanehisa Laboratories


Glioma





Stratification based on clinical variable (gender)

How to Choose Stratifications?

~ 15 clusterings per matrix

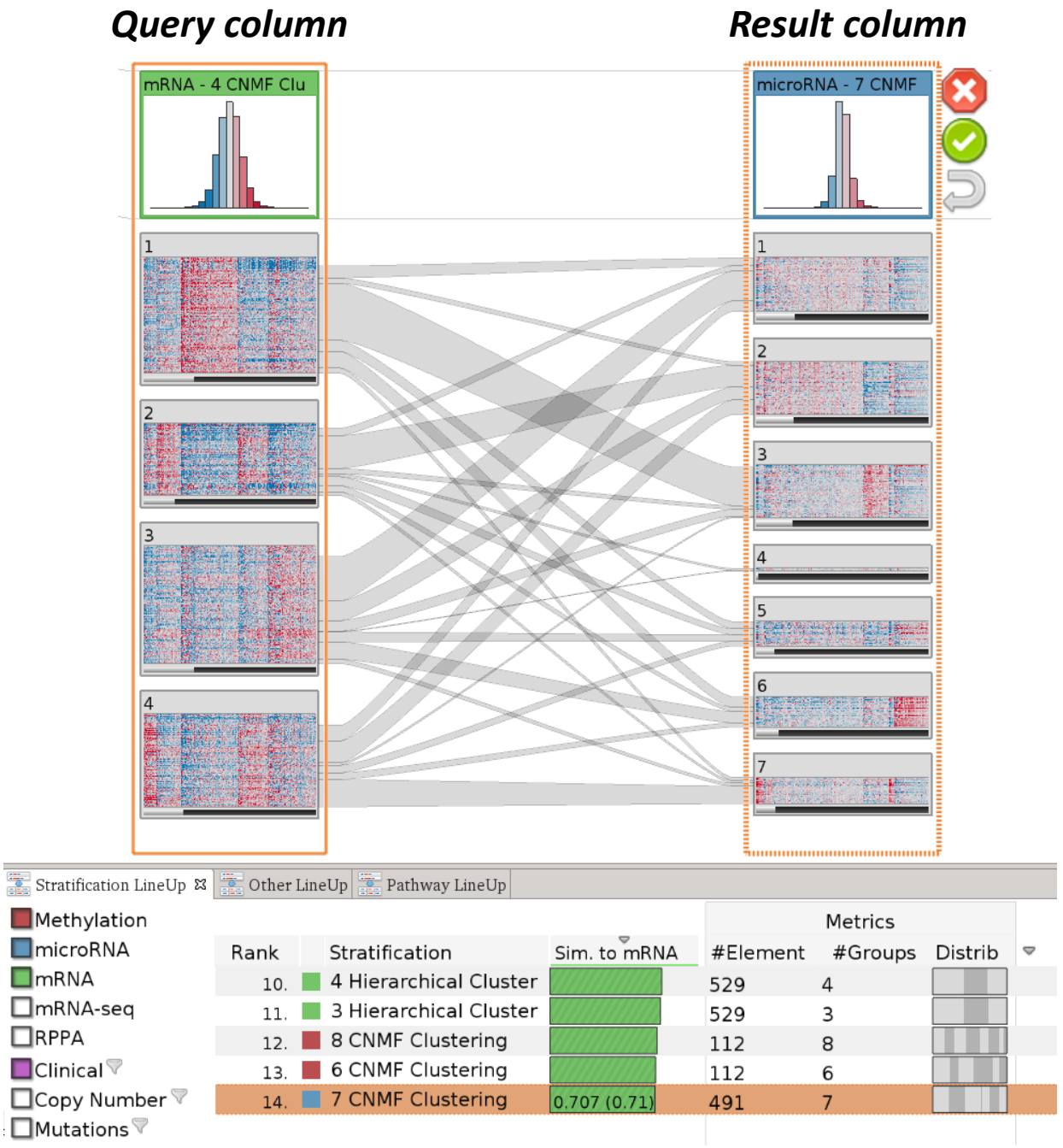
~ 15,000 stratifications for copy number & mutations

~ 500 pathways

~ 20 clinical variables

Calculating scores for matches

Ranking the results



Considered Datasets

Ranked Stratifications

Algorithms for finding..

… matching stratification

… matching subtype

… mutual exclusivity

… relevant pathway

… stratification with significant effect in survival

… high/low structural variation

Live-Demo!

http://stratomex.caleydo.org

Caleydo enRoute

PATHWAYS & EXPERIMENTAL DATA

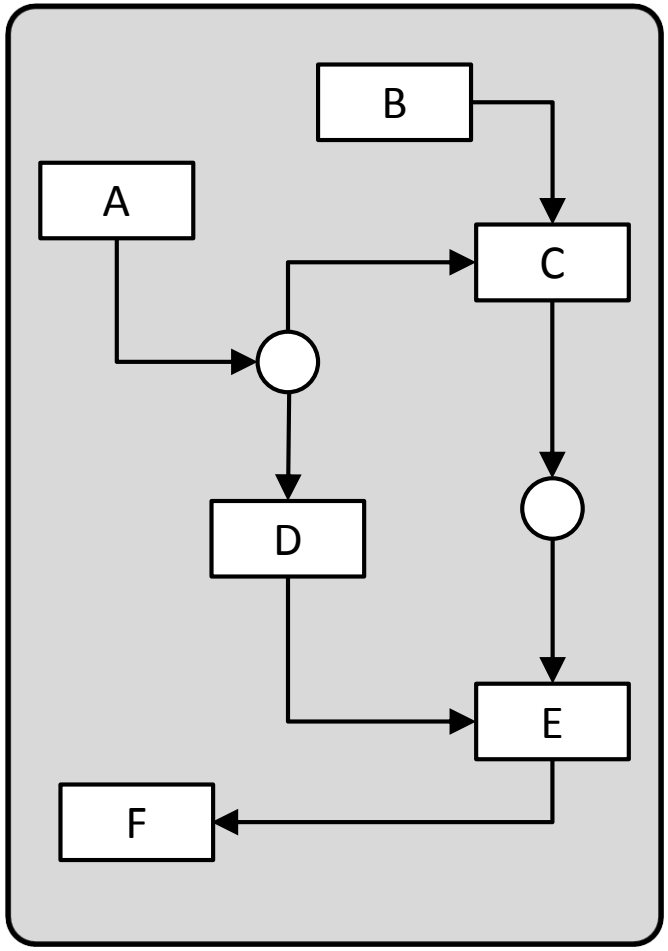
Experimental Data and Pathways

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.

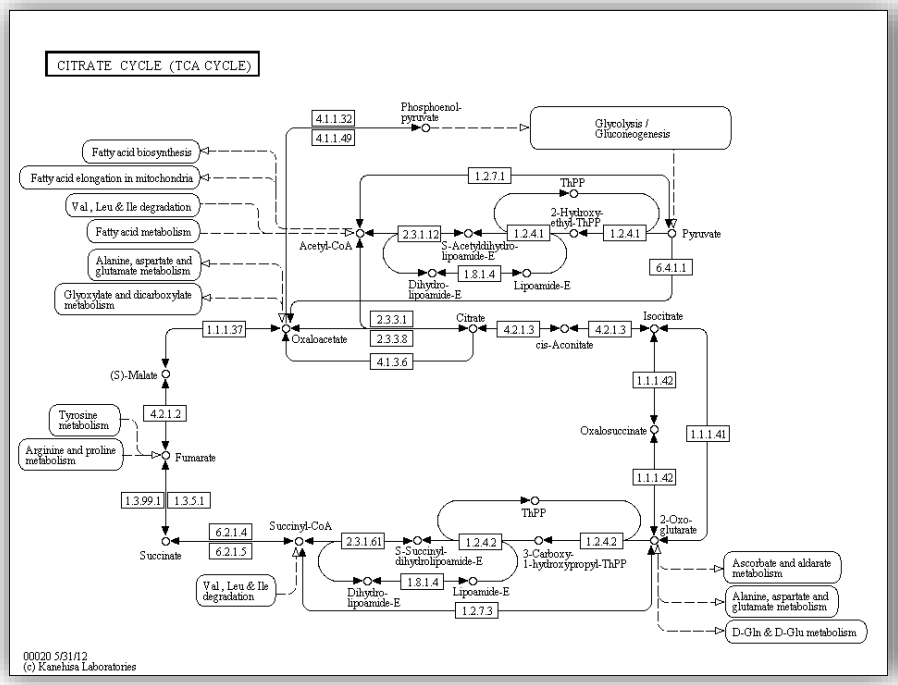
Why use Visualization?



Efficient communication of information

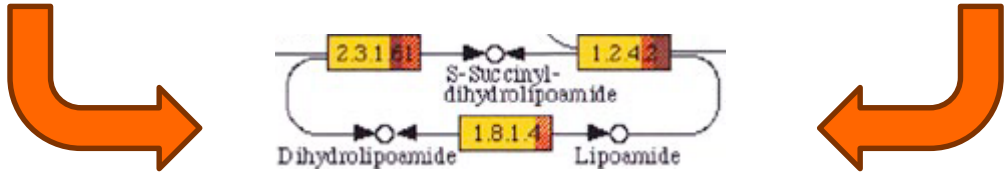
A	-3.4
B	2.8
C	3.1
D	-3
E	0.5
F	0.3

Experimental Data and Pathways



	A	B	C	D	E	F	G
1	Hybridization	0001-01C-01R-0177-01	0003-01A-01R-0177-01	0004-01A-01R-0298-01	0007-01A-01R-0177-01	0009-01A-01R-0177-01	0010-01A-01R-0177-01
2	Composite Ele	Signal	Signal	Signal	Signal	Signal	Signal
3	AACS	0.51414972642765	0.408345300354386	103.839.280.872.892	108.369.722.813.568	146.228.041.324.369	189.518.994.867.335
4	FSTL1	274.326.077.373.904	366.350.031.165.199	572.009.498.050.765	-115.814.163.115.937	462.733.779.632.644	283.075.033.965.805
5	ELMO2	-0.47503962404922	0.083081303877976	0.71228305118261	-0.872750104404229	0.407918562894304	145.159.574.522.917
6	CREB3L1	-110.344.925.188.626	-129.462.362.872.386	-0.227010846693110	-0.284588481211029	-0.879544579484095	-0.0134788669146966
7	RPS11	499.838.216.564.566	468.034.485.053.486	409.691.860.929.175	437.410.995.409.265	442.451.017.149.714	490.071.296.665.235
8	PNMA1	154.879.089.253.498	0.867033048134666	27.926.044.547.714	225.292.305.165.001	0.730144175027465	194.355.497.056.736
9	MMP2	268.760.812.835.858	321.769.033.459.614	490.900.567.411.305	-167.367.360.374.371	323.947.380.237.154	-159.909.546.919.508
10	SAMD4A	-0.95385000100184	-110.393.925.981.711	-0.780339641097821	-0.927255839038269	-104.784.345.382.755	-113.068.517.857.268
11	SMARCD3	-12.754.315.263.248	-0.803166248235193	192.878.200.580.422	-166.272.411.648.149	-126.054.982.166.486	253.814.516.975.109
12	A4GNT	-0.8779238907219	-178.229.413.280.745	-22.870.781.160.139	-127.901.084.268.365	-200.815.270.464.235	-137.113.826.496.947
13	C9orf39	-119.755.123.311.627	-199.689.453.294.211	-170.879.710.400.263	-197.903.696.598.943	-184.196.451.349.235	-123.143.505.917.242
14	PKNOX2	-0.95454191801256	-129.980.559.337.036	-129.725.303.429.725	-117.863.280.467.411	-169.952.892.104.581	-0.914002701548987
15	RALYL	-163.237.829.190.046	-223.568.546.117.929	-202.107.025.240.506	-167.211.920.052.532	-226.223.976.628.069	-180.973.590.502.948
16	ZHX3	-0.92856964251013	-111.244.404.270.456	0.0565663069364497	-0.794043792350179	-101.938.096.745.295	-0.413263822075167
17	ERCC5	0.95243371908686	0.957680305439156	0.99553222244338	0.238642819120941	153.006.779.612.285	-0.071226896339307
18	RXFP3	-206.267.350.950.005	-247.628.615.578.476	-258.794.340.437.298	-243.076.988.610.523	-261.197.965.543.047	-216.635.887.704.531
19	APBB2	-101.794.466.906.032	-139.226.919.342.780	0.7396843984963	-160.525.972.599.955	-129.292.861.167.009	-153.166.942.410.299
20	BBOX1	-219.765.356.842.015	-216.619.193.833.608	189.217.944.622.057	-223.116.920.533.355	-242.172.086.385.896	-172.405.872.547.507
21	PRO0478	-208.735.486.814.708	-201.266.491.273.900	-264.514.012.419.739	-183.469.408.961.363	-228.483.271.853.159	-222.688.943.536.626
22	GCSH	214.667.175.620.209	214.425.000.089.483	282.468.317.897.215	379.400.938.009.765	234.157.818.929.887	305.821.981.214.471
23	XDH	-140.335.573.729.949	-154.352.054.860.075	-195.598.156.263.592	-164.927.258.900.381	-153.786.759.540.717	-150.754.123.452.277
24	EDN1	-0.27265835556173	0.0601082045242967	-0.686221828128451	-0.070390294616588	0.241048141573394	-0.681241448310897
25	MTERF	0.33456496873088	0.602787120981156	116.205.672.714.072	0.290137408663012	0.339834697015724	0.376698230530083
26	PDCL3	0.22161805747679	192.980.867.583.586	171.282.568.732.192	129.218.742.469.691	183.026.664.772.145	211.448.961.587.135
27	CLK4	0.11771672245876	0.837990640842366	0.48153932013633	0.101190616384591	0.461222798318555	0.547200380309174
28	KCNG1	-145.735.076.947.667	-154.967.690.506.818	-173.914.329.989.365	-168.541.602.431.989	-196.875.877.389.228	-109.608.188.831.993
29	CXCR4	187.536.347.744.716	263.335.862.666.286	228.888.923.555.987	-145.426.654.150.337	277.993.551.197.589	32.964.742.520.002
30	DECR1	267.353.839.615.507	243.484.300.334.252	340.595.900.405.511	332.766.227.083.077	300.755.122.393.034	195.692.044.856.851
31	SALL1	-131.231.146.088.135	-206.830.763.066.117	-167.501.082.187.998	-196.742.986.889.255	-226.876.961.736.839	-0.851620028342527
32	PTPRR	-175.072.128.842.408	-215.050.278.107.699	-201.907.311.040.685	-0.618976507439188	-0.453120032708616	166.423.598.869.425
33	CADM4	-159.091.940.638.954	-196.549.713.788.350	-224.919.487.589.083	-163.188.842.517.916	-202.074.958.245.880	-179.879.615.356.428
34	IRAK1	315.954.017.145.736	299.254.943.418.361	356.221.429.575.433	368.207.420.675.573	332.140.514.206.449	197.043.270.833.661

[KEGG]



[Lindroos2002]

Five Requirements

Ideal visualization technique addresses all

Talking about 3 today

R 1: Data Scale

Large number of experiments

- Large datasets have more than 500 experiments

Multiple groups/conditions

R II: Data Heterogeneity

Different types of data, e.g.,

mRNA expression	*numerical*
mutation status	*categorical*
copy number variation	*ordered categorical*
metabolite concentration	*numerical*

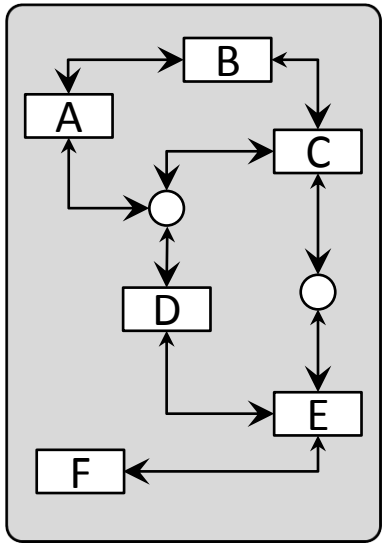
Require different visualization techniques

R V: Supporting Multiple Tasks

Two central tasks:

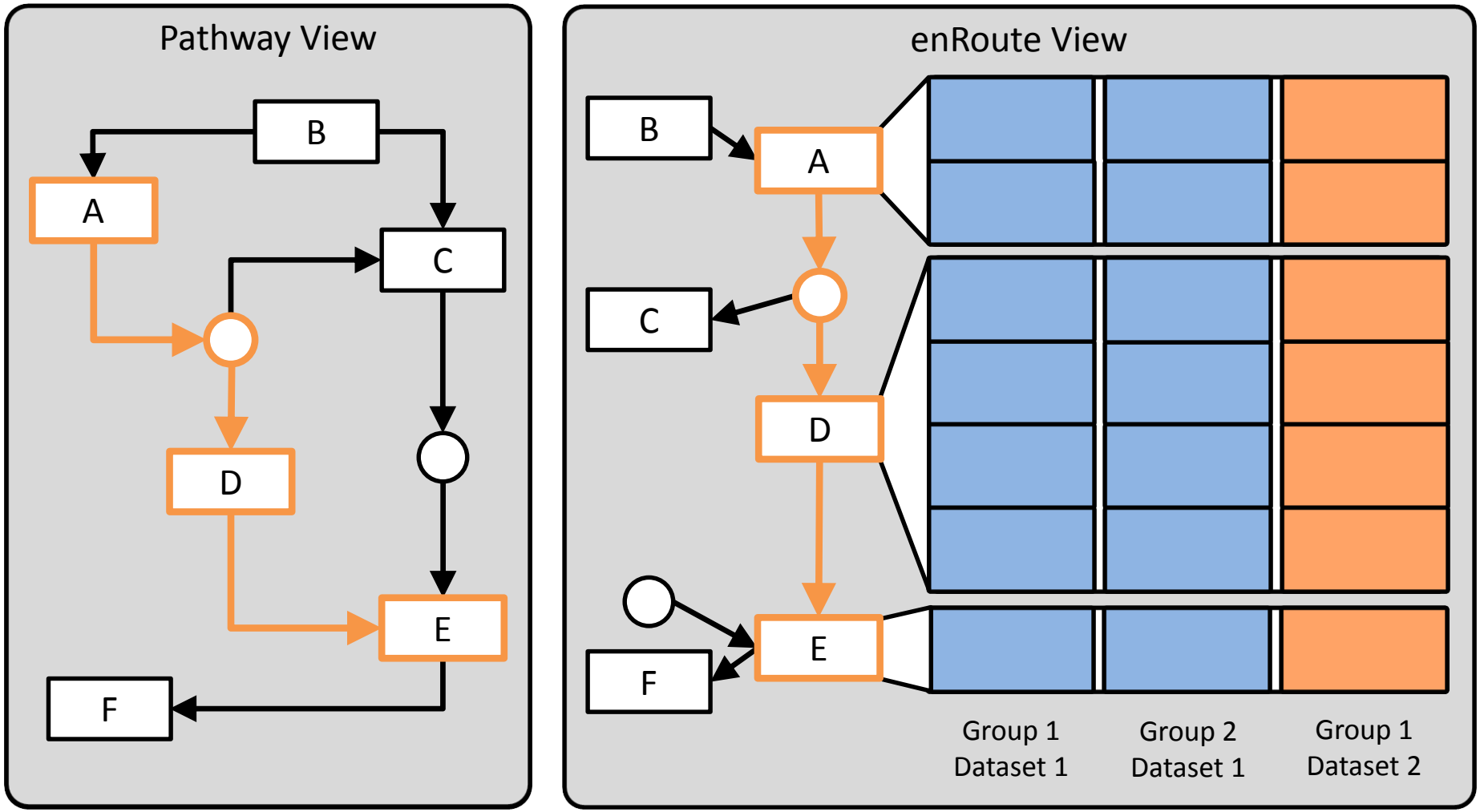
- Explore topology of pathway
- Explore the attributes of the nodes (experimental data)

Need to support both!

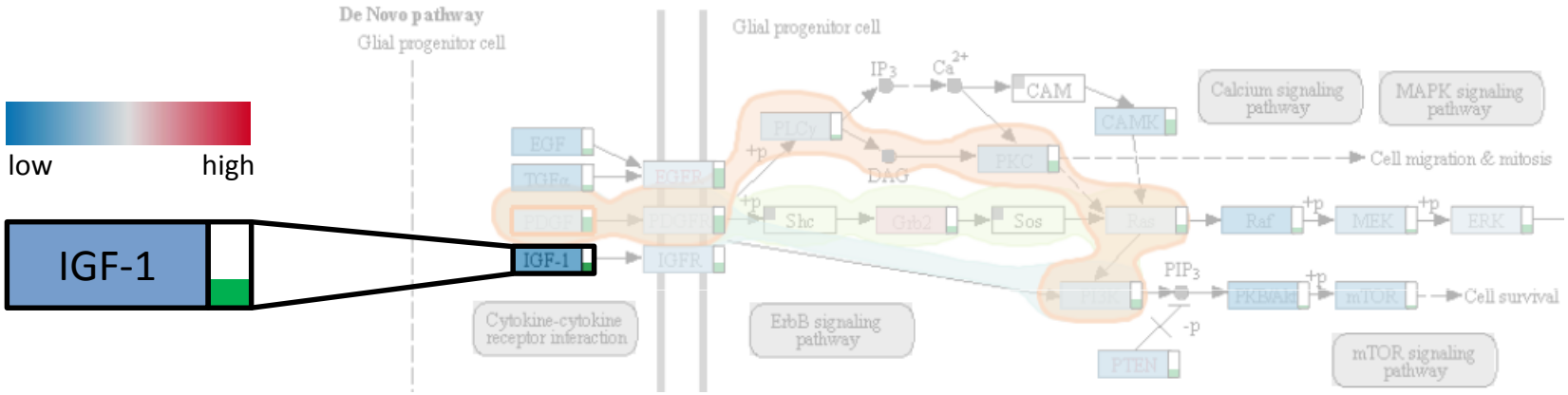


	Sample 1	Sample 2	Sample 3
Gene 1	1	1.1	0.4
Gene 2	2	0.5	1.2
Gene 3	1.4	0.2	0.5
Gene 4	0.3	0.5	0.7

Concept


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

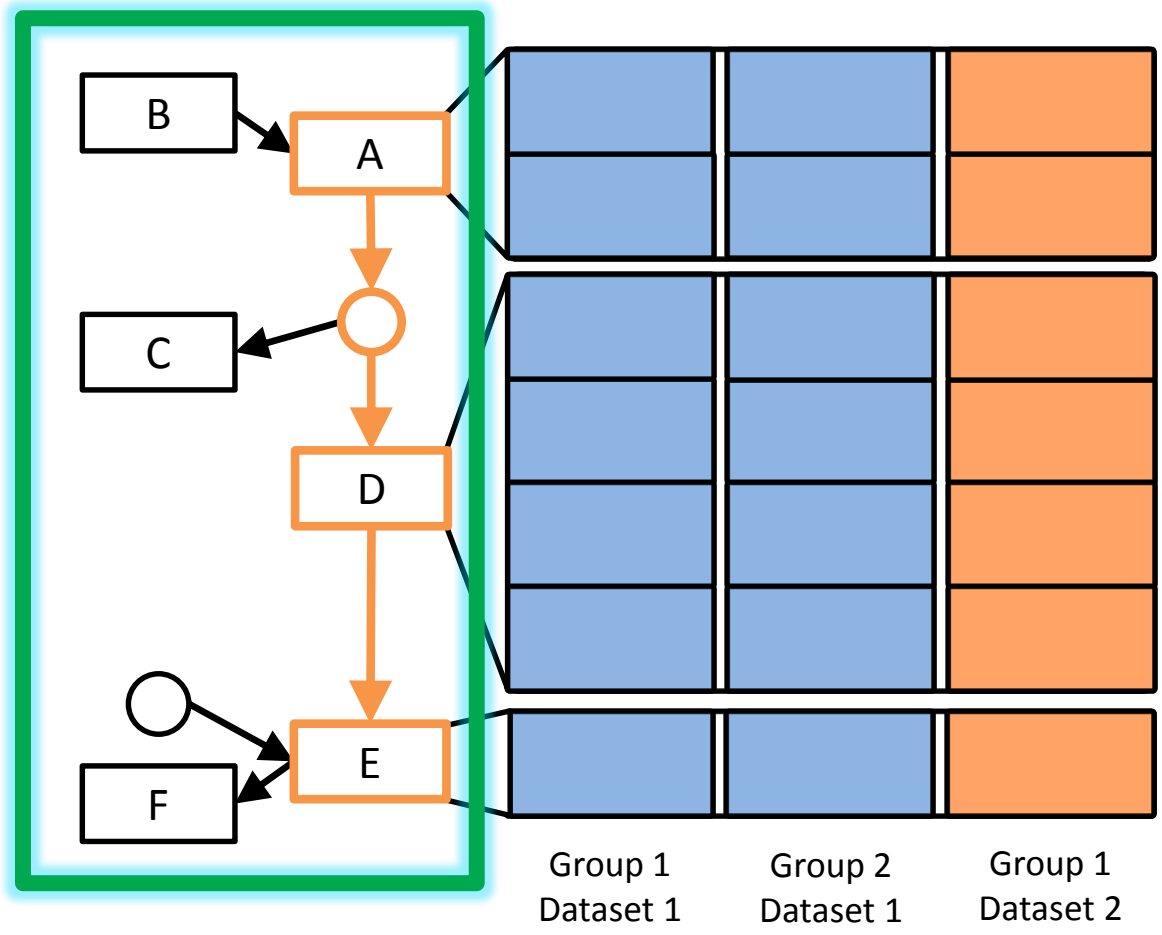
Pathway View



On-Node Mapping

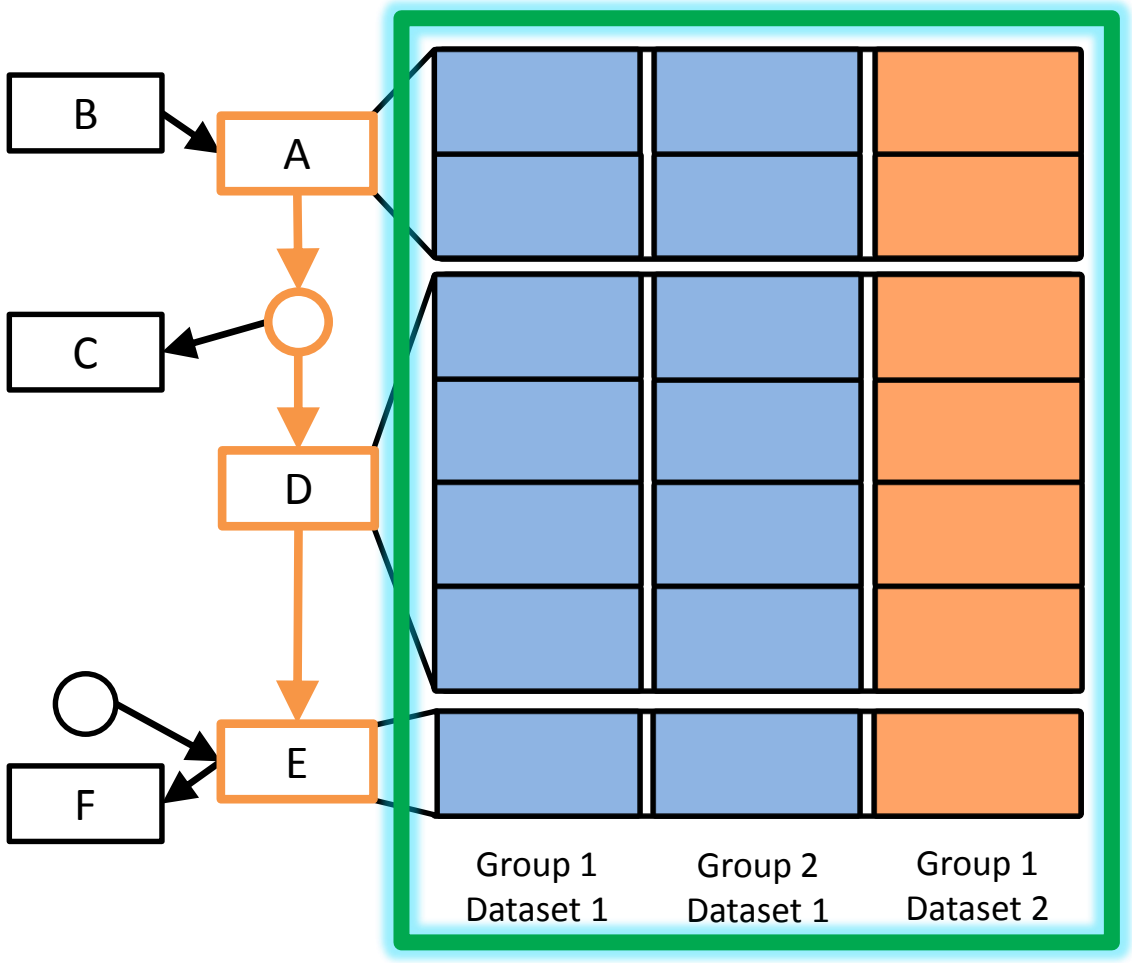
Path highlighting with Bubble Sets [Collins2009]

enRoute View



Path Representation

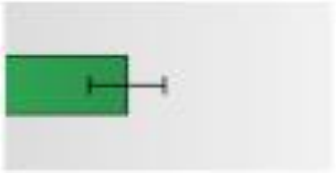
enRoute View



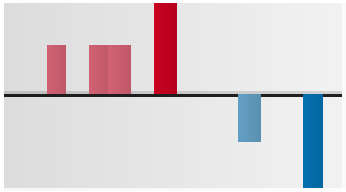
Experimental Data Representation

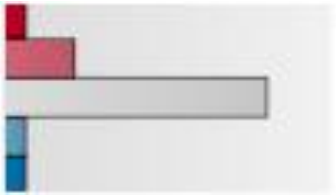
Experimental Data Representation

Gene Expression Data (Numerical)

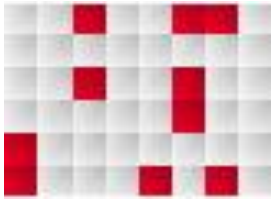


Copy Number Data (Ordered Categorical)

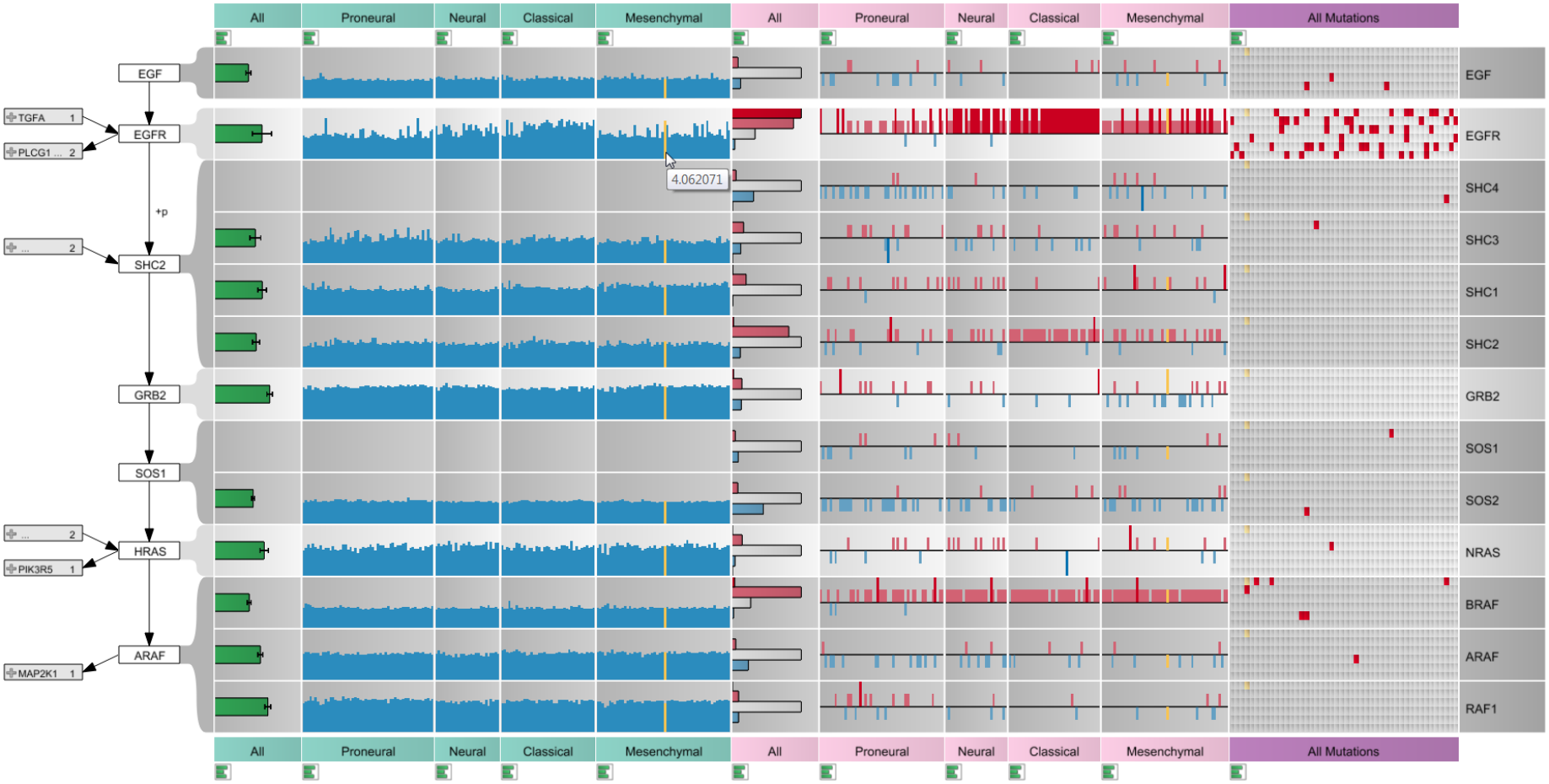




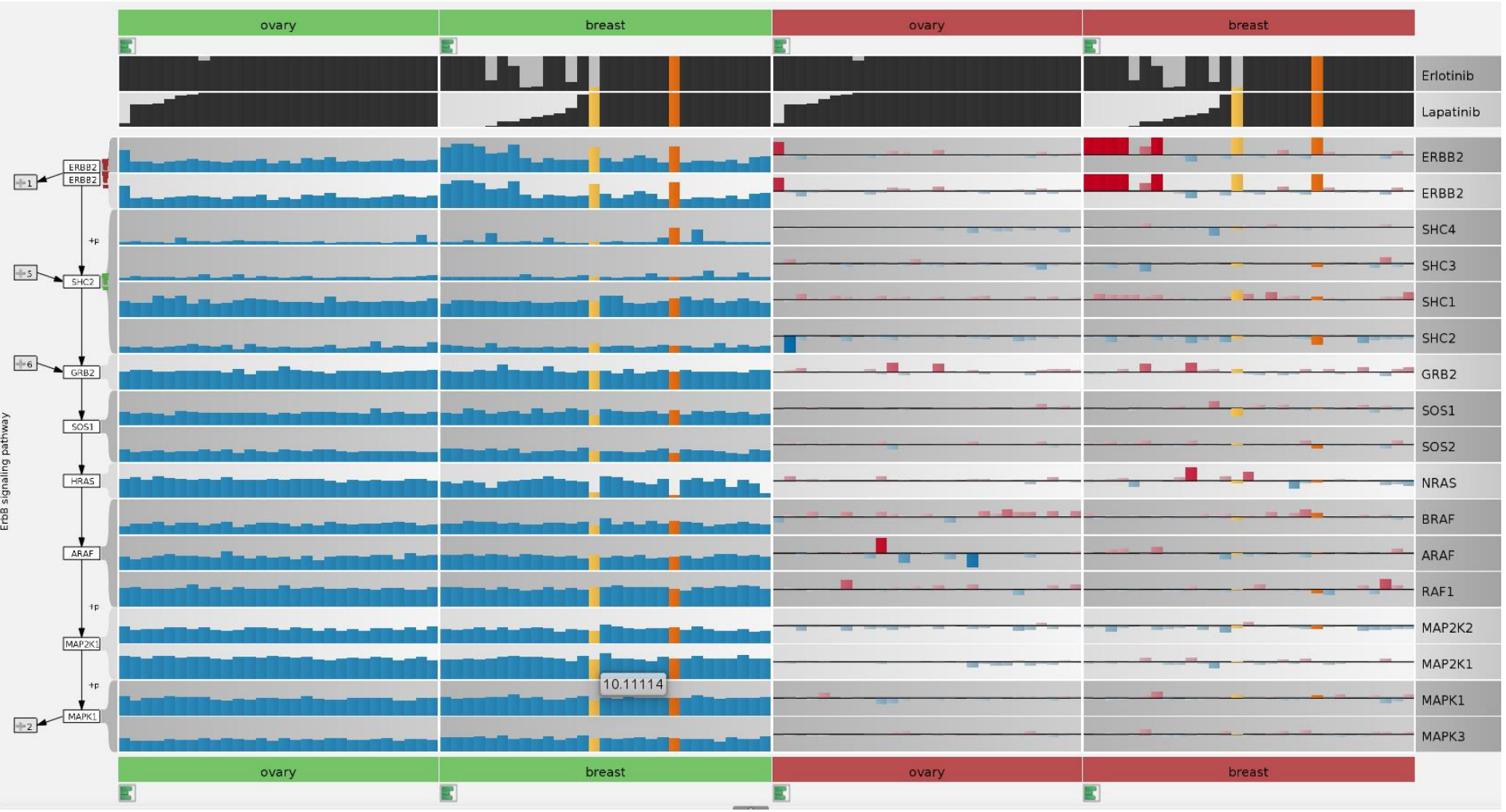
Mutation Data



enRoute View – Putting All Together



CCLE Cell lines & Cancer Drugs



Analysis by Anne Mai Wasserman

Collaboration with AM Wassermann, M Borowsky, M Glick @NIBR

MANAGING PATHWAYS & CROSS-PATHWAY ANALYSIS

Pathways

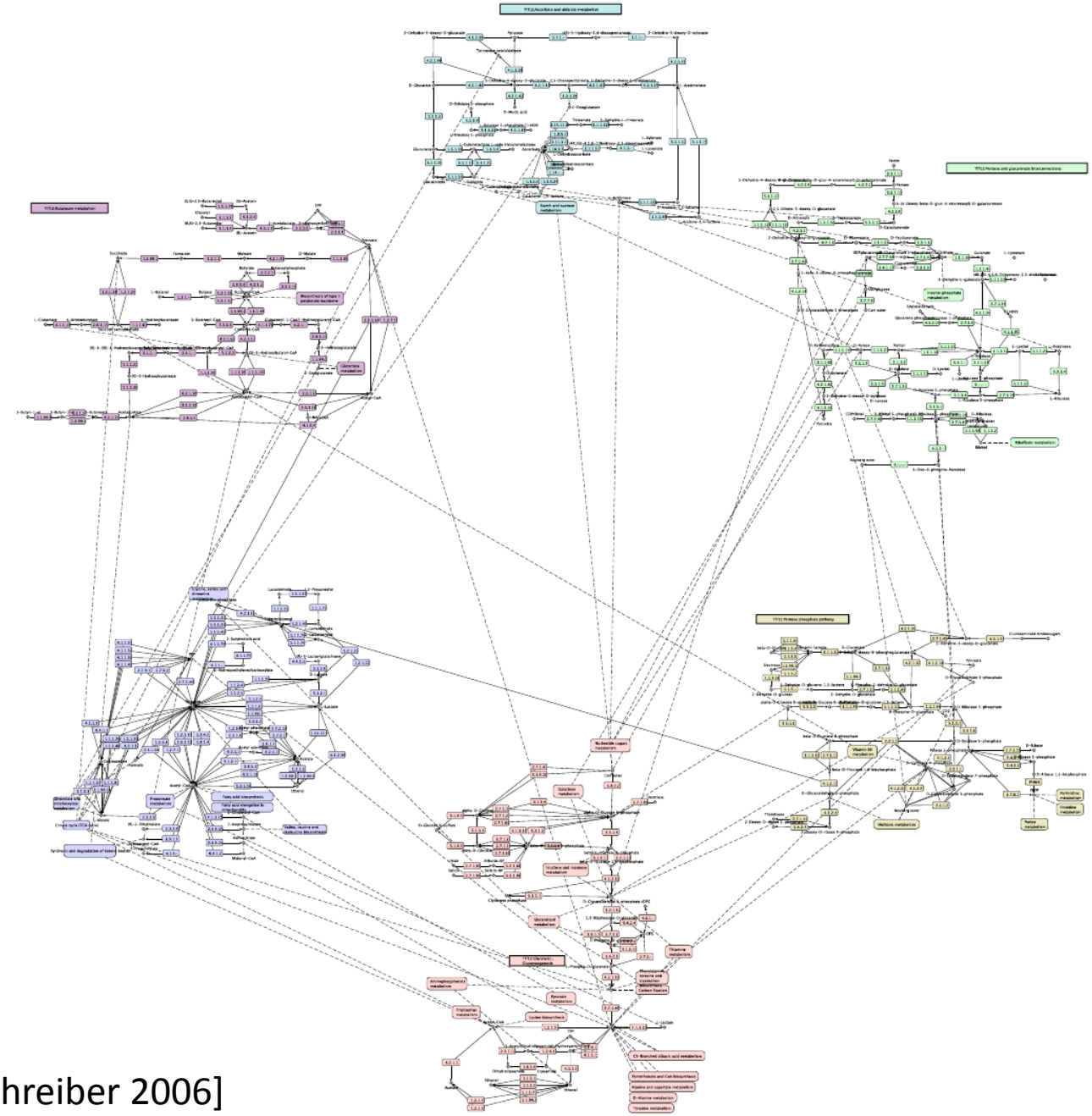
Partitioning in pathways is *artificial*

Purpose: reduce complexity

- "Relevant" subset of nodes and edge

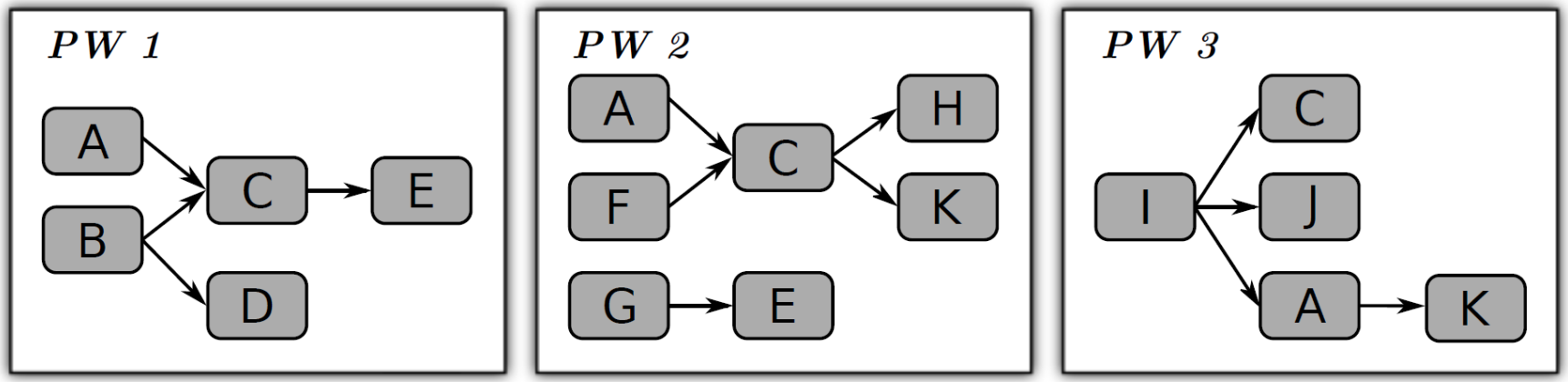
Makes it hard to

- understand cross-talk
- identify role of nodes in other pathways

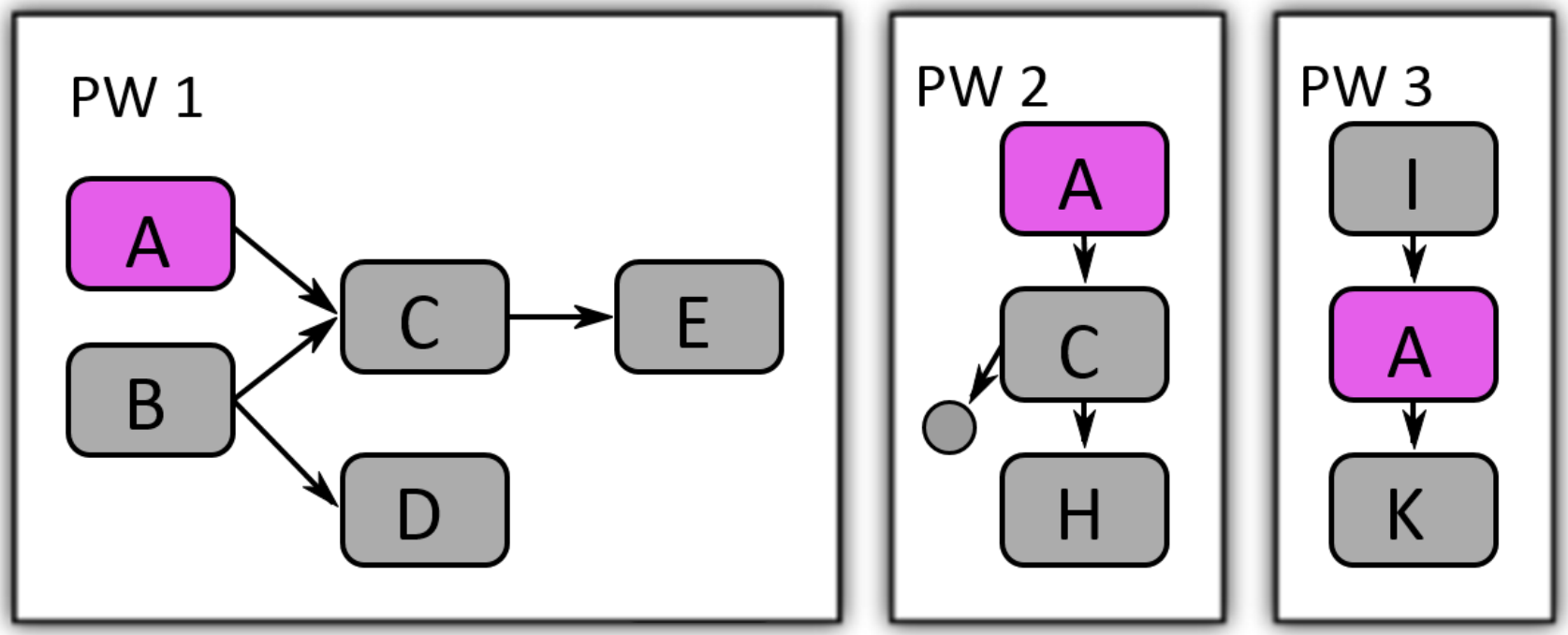


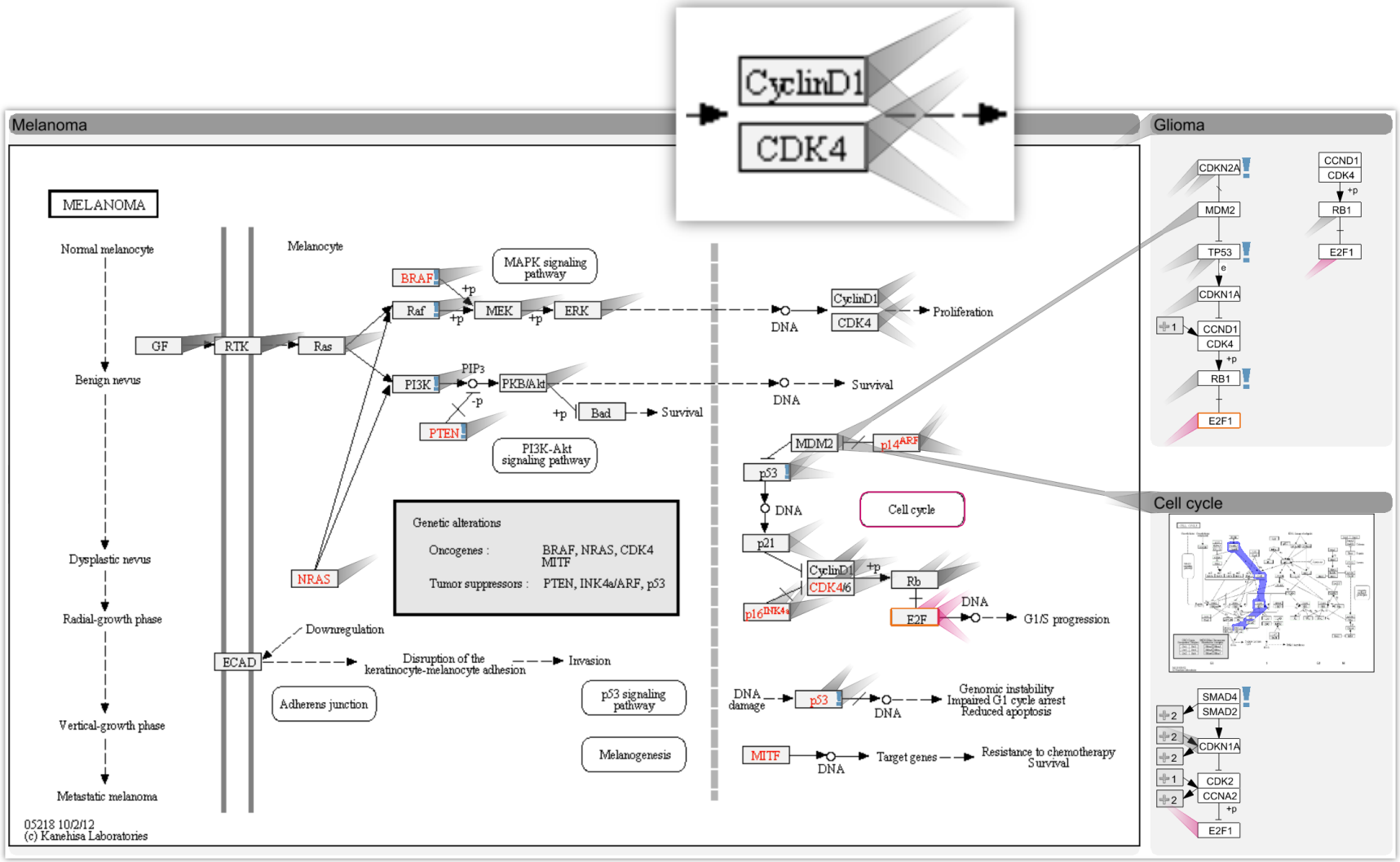
[Klukas & Schreiber 2006]

Solution: Contextual Subsets



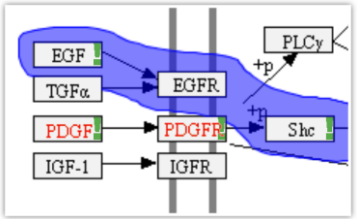
Solution: Contextual Subsets

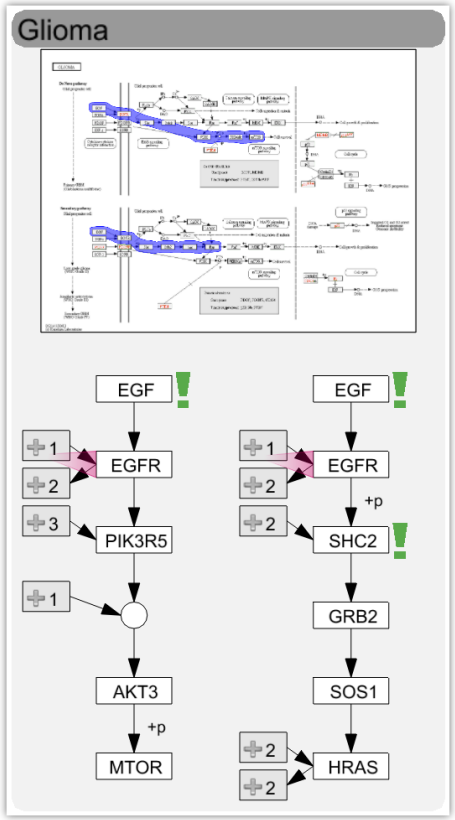


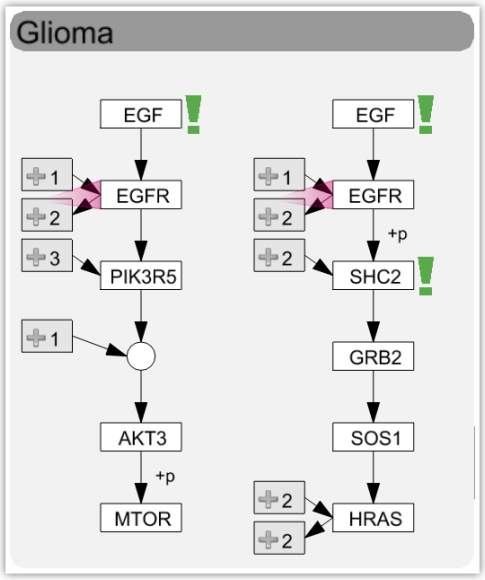

CyclinD1
CDK4
Melanoma
MELANOMA
Normal melanocyte
Melanocyte
Benign nevus
Dysplastic nevus
Radial-growth phase
Vertical-growth phase
Metastatic melanoma
GF
RTK
Ras
BRAF
Raf
MEK
ERK
MAPK signaling pathway
PI3K
PIP3
PKB/Akt
Bad
Survival
PTEN
PI3K-Akt signaling pathway
NRAS
Genetic alterations
Oncogenes : BRAF, NRAS, CDK4 MITF
Tumor suppressors : PTEN, INK4a/ARF, p53
ECAD
Downregulation
Disruption of the keratinocyte-melanocyte adhesion
Invasion
Adherens junction
p53 signaling pathway
Melanogenesis
DNA
CyclinD1
CDK4
Proliferation
MDM2
p14ARF
p53
p21
Cell cycle
CyclinD1
CDK4/6
p16INK4a
Rb
E2F
G1/S progression
DNA damage
Genomic instability
Impaired G1 cycle arrest
Reduced apoptosis
MITF
Target genes
Resistance to chemotherapy
Survival
05218 10/2/12
(c) Kanehisa Laboratories
Glioma
CDKN2A
MDM2
TP53
CDKN1A
CCND1
CDK4
RB1
E2F1
CCND1
CDK4
RB1
E2F1
Cell cycle
SMAD4
SMAD2
CDKN1A
CDK2
CCNA2
E2F1

Levels of Detail

Thumbnail showing paths







Experimental data highlights

How to Select Pathways?

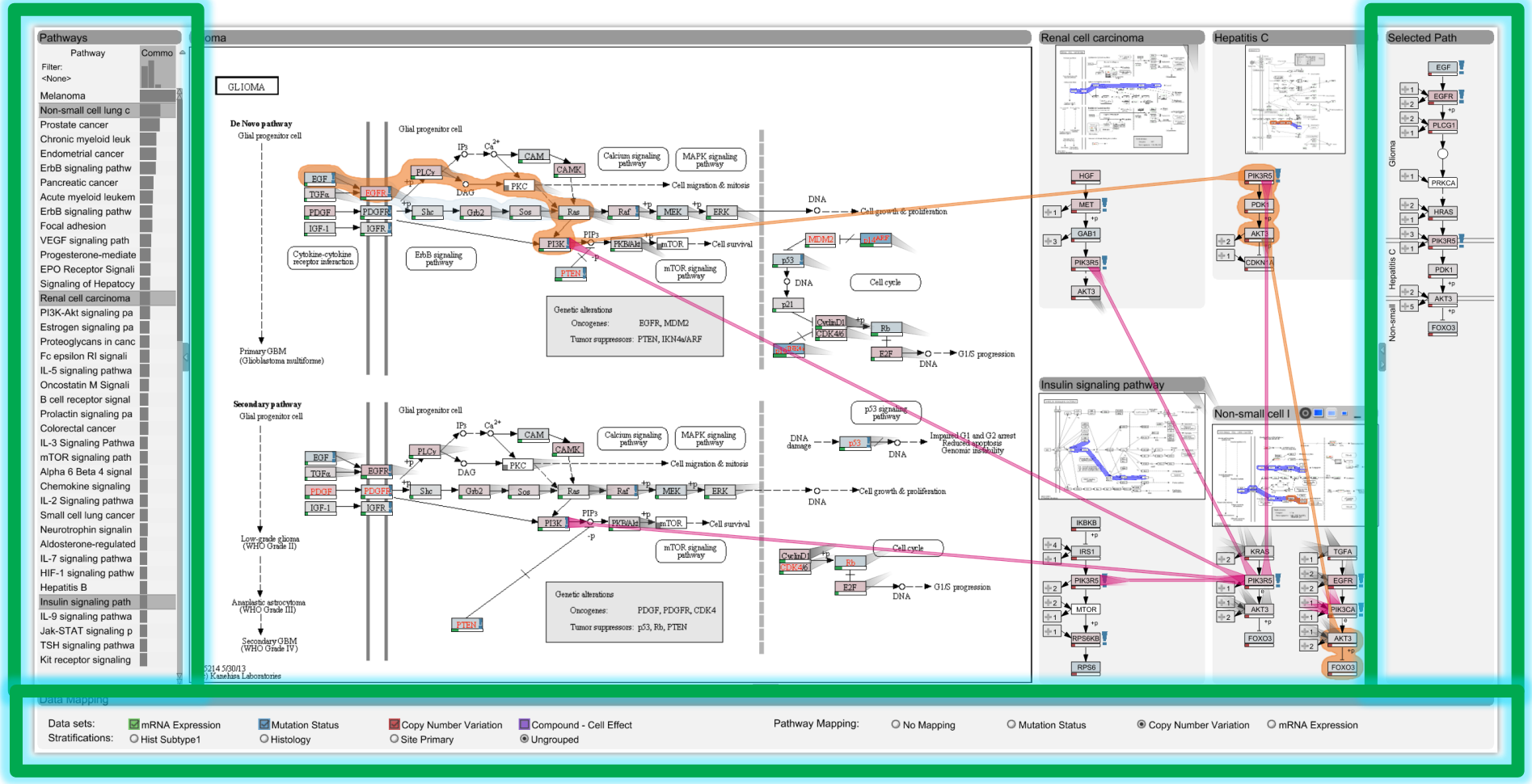
Search Pathway

Find pathways that contain *focus node*

Find pathway that is similar to another one

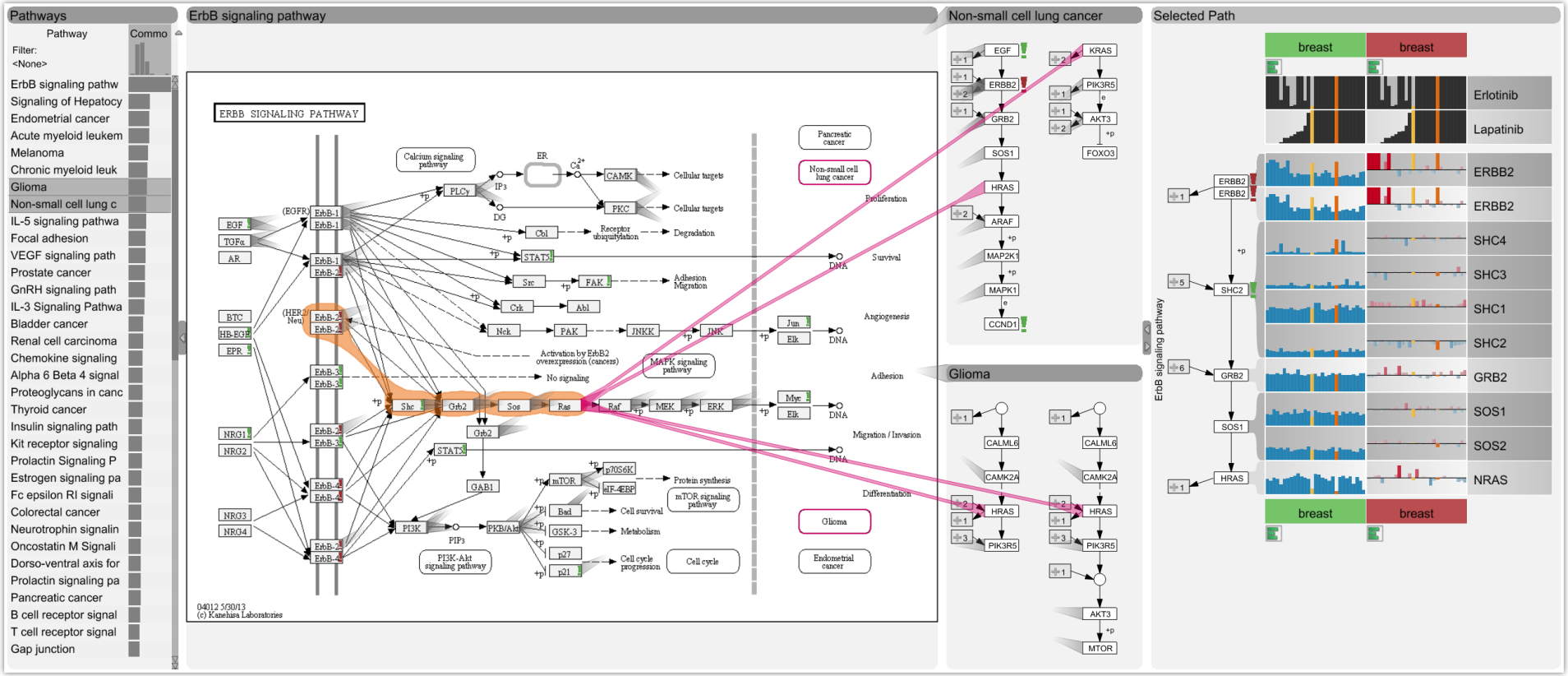
Ranked pathway list

Selected Path



Datasets

Integrating enRoute



Video!

http://enroute.caleydo.org

More Information

http://caleydo.org

Software, Help, Project Information, Publications, Videos

Data Visualization In Molecular Biology

?

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

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