Genomic Visualization

Mahika Dubey, Hadiseh Gooran, Veronica Rivera
CMPM 290A Immersive Analytics

Daniel Kim - Kim Lab

(https://dkim.sites.ucsc.edu)

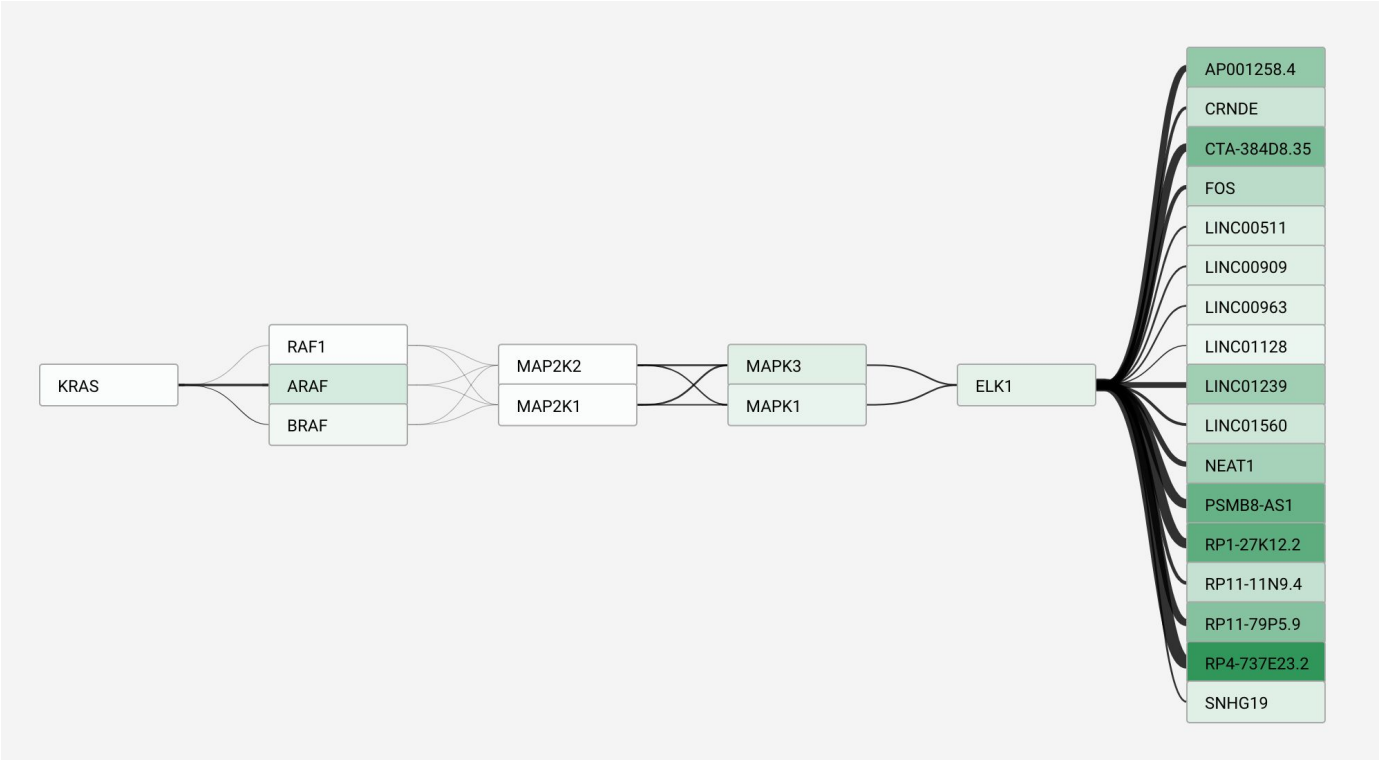
- Study noncoding RNA in stem cells and cancer
- Use genomic technologies to characterize noncoding RNA in single cells
- kRas: driver mutation--one of the earliest mutations that causes cells to become cancerous
- Lab does a lot of sequencing and generates sequencing data
- Trying to understand different pathways between proteins and RNA sequences, and how these contribute to cancer

Angela Brooks - Brooks Lab

(https://brookslab.soe.ucsc.edu)

- Focus on studying somatic mutations that cause changes to the transcriptome
- Trying to understand how alternative splicing is regulated
- Develop computational techniques to analyze genome and transcriptome sequencing data

Existing Visualization



Created for a publication using **D3** by undergraduate researcher **Cassia Artanegara** in collaboration with Kim Lab

Insights from PI

- 2 main goals for visualizations
  - Static diagrams for publications
  - Dynamic & Interactive visualizations for data exploration within research groups

Current Ideas

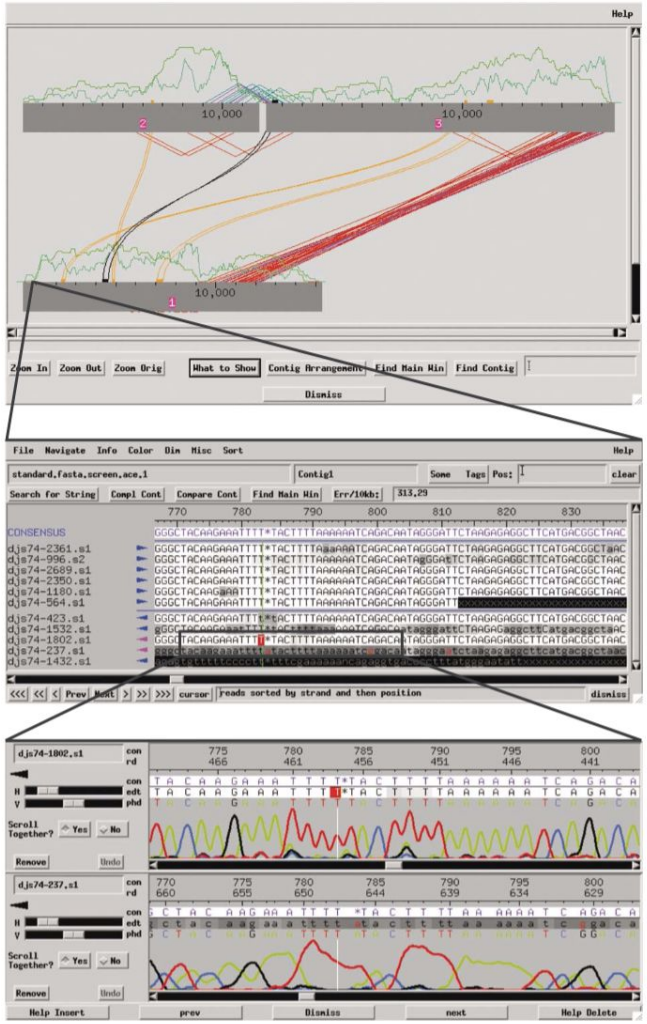
- **Main goal: Pathway visualization that helps visualize a signaling cascade; building on RasVis**
- Automated generation of network based on some user provided data files, rather than hard coded data
- Tool that parses data into a common format for visualizing
- Hyperlinks and metadata on hover
- Take group of genes, cross reference with sequencing data and visualize the hierarchies.
- Ability to add new data that forms new pathways from existing ones
- **Long term:** take genes from data input, find relevant pathways, and visualize

Visualizing genomes: techniques and challenges

data analysis is replacing data generation as the rate-limiting step in genomics studies.

Genomic data visualization tools for **researchers** to:

- Explore the data
- Interpret the data
- manipulate the data



‘Assembly view’

as gray boxes with a scale of nucleotide positions within the contig.

- **Angled colored lines** represent **mate pairs**
- **Curved lines** indicate **sequence similarity computed using cross_match**
- **Read coverage** is plotted along the **contigs in dark green mate-pair coverage highlighted in light green.**

‘Aligned reads’

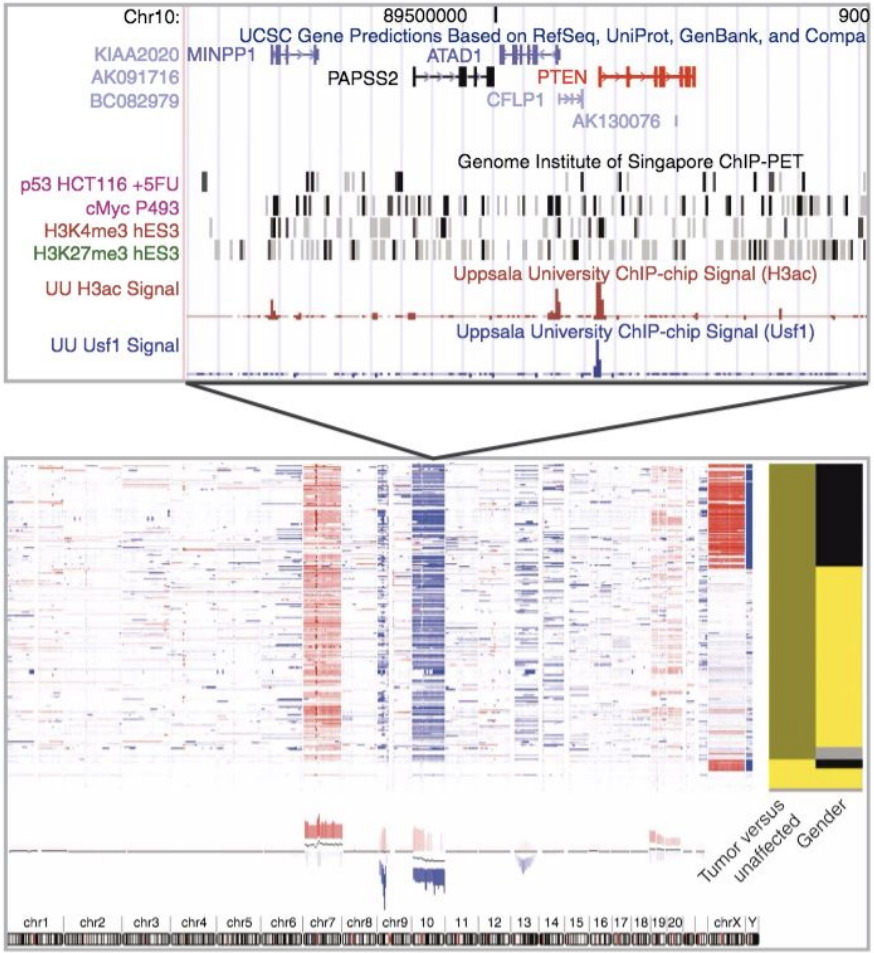
displays a vertical stack of read sequences, optionally separated by strand, with forward on top (right arrows) and reverse on bottom (left arrows).

The ***** character in the computed consensus indicates that one or more of the reads contains an insertion at this position that the assembly program deems incorrect.

‘Trace’

Users can:

- **evaluate the insertion**
- **override the assembly program’s choice of consensus**

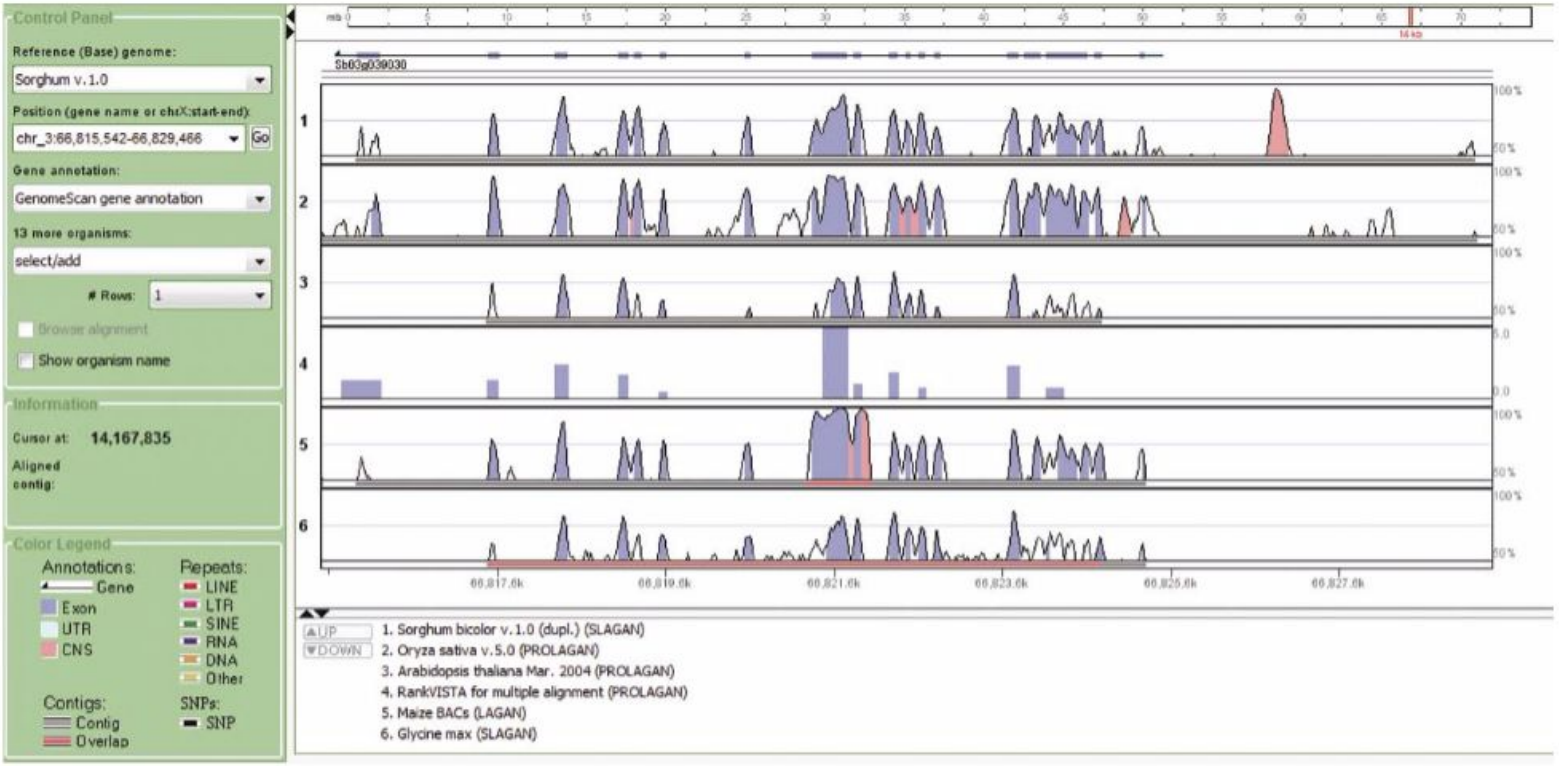


The UCSC Genome and Cancer Genomics Browsers

Displays diverse data types across the human reference assembly (for example, gene annotations with exons (boxes), introns (thin lines) and untranslated regions (intermediate-height boxes); ChIP data as heat maps or histograms)

The UCSC Cancer Genomics Browser provides

an improved overview and links back to the Genome Browser. Two publicly available clinical parameters are displayed:
tumor (olive) versus **unaffected** (yellow)
male (yellow) versus **female** (black); gray, data unavailable



VISTA

This plot corresponds to a 14-kb interval on the Sorghum bicolor v.1.0 assembly (chr. 3, 66815542–66829466).
Conserved regions are colored according to the gene annotation displayed above the conservation plot (blue, conservation in exons; light blue, in untranslated regions; pink, in conserved noncoding sequences).
Several alignments can be viewed at the same time to assist in analysis.

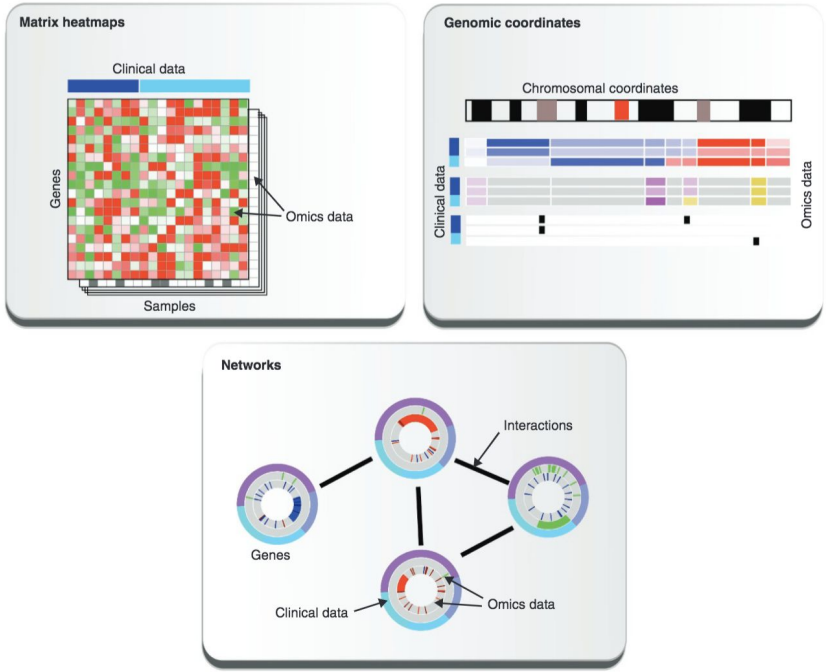
Visualizing multidimensional cancer genomics data

the use of intuitive visualization tools : **cancer genomics**

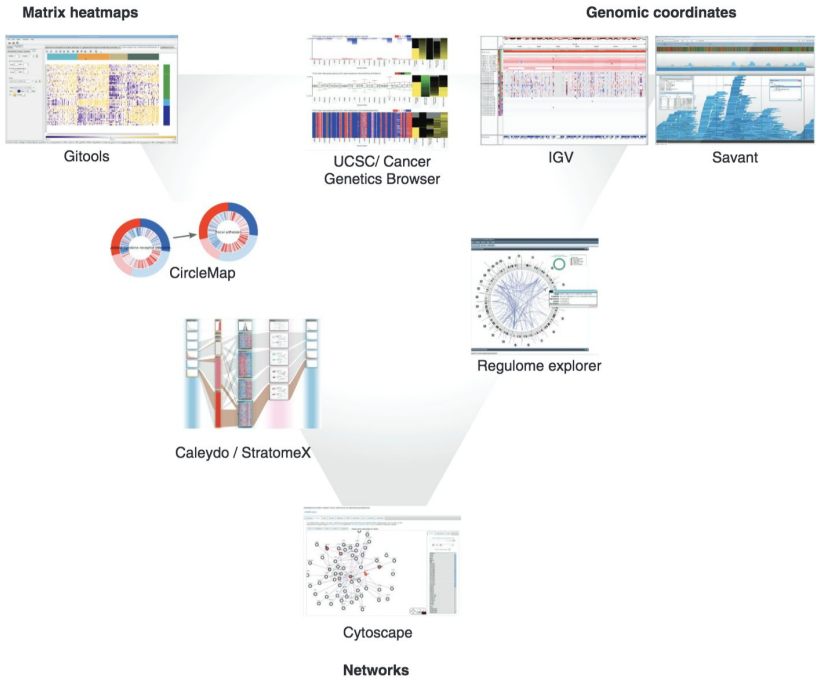
Effective and common visualization techniques

- exploring oncogenomics data
- discuss a selection of tools that allow researchers to effectively visualize multidimensional oncogenomics datasets

Tools such as:
Circos, Gitools, the Integrative Genomics Viewer, Cytoscape, Savant Genome Browser, StratomeX and platforms such as cBio Cancer Genomics Portal, IntOGen, **the UCSC Cancer Genomics Browser**, the Regulome Explorer and the Cancer Genome Workbench.

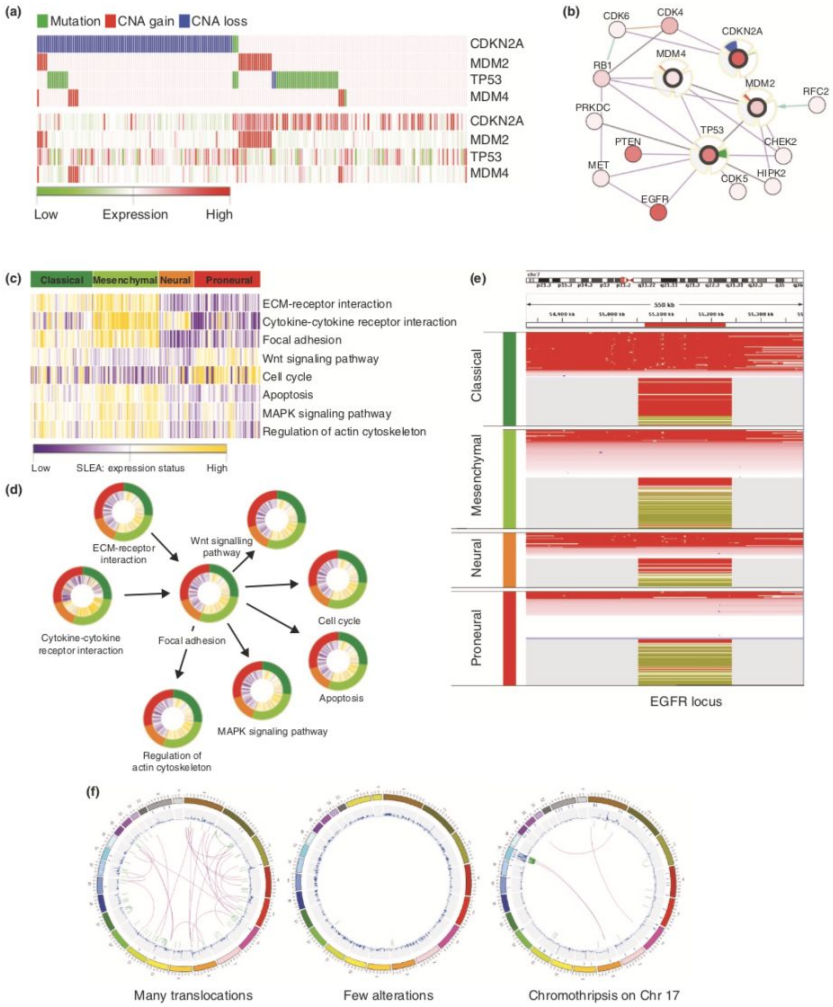


Cancer genomics projects generate multidimensional data for a cohort of patients



Screenshots of tools that are frequently used in cancer genomics research distributed according to their visualization principles

.



Cancer genomics projects generate multidimensional data for a cohort of patients

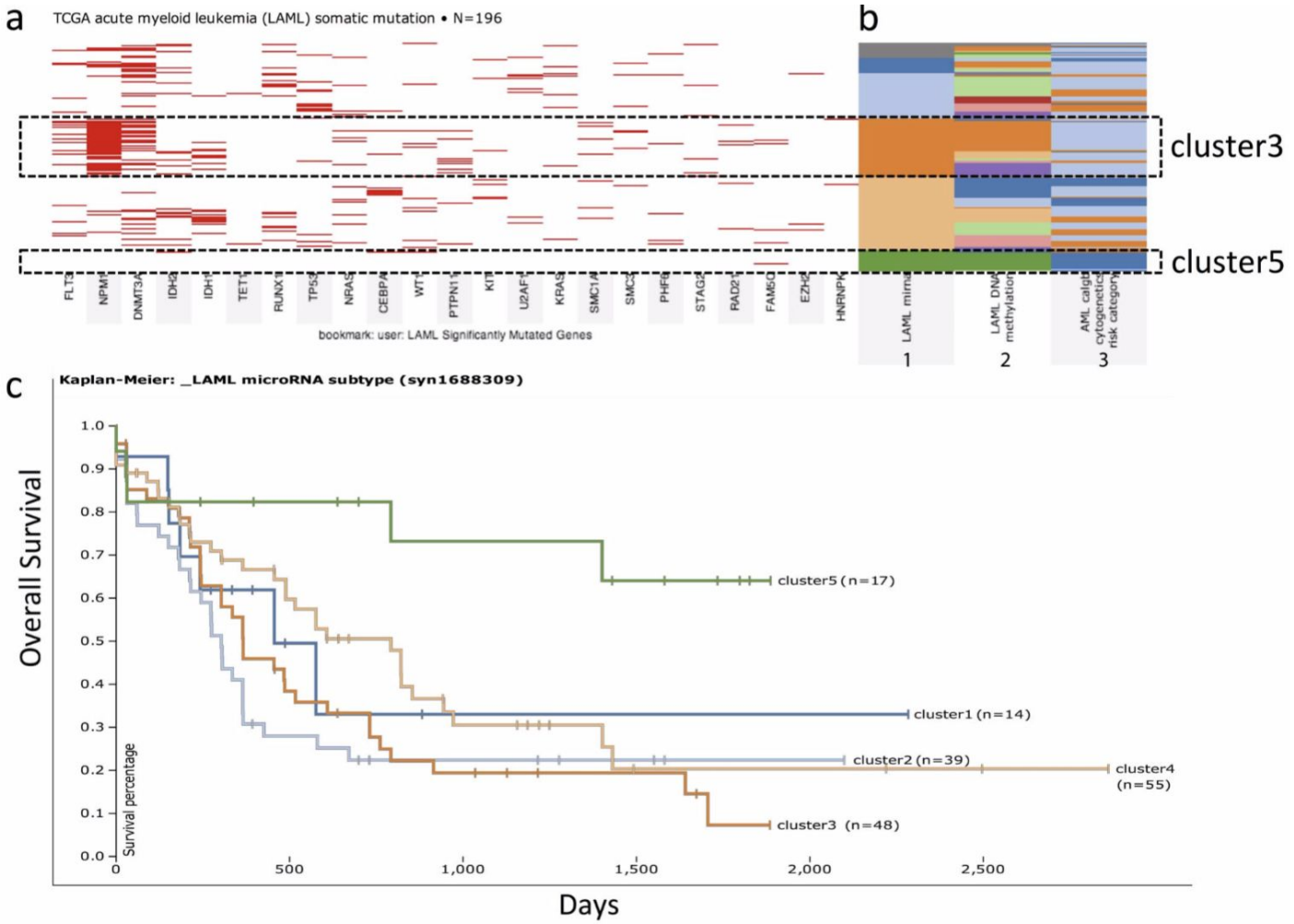
.

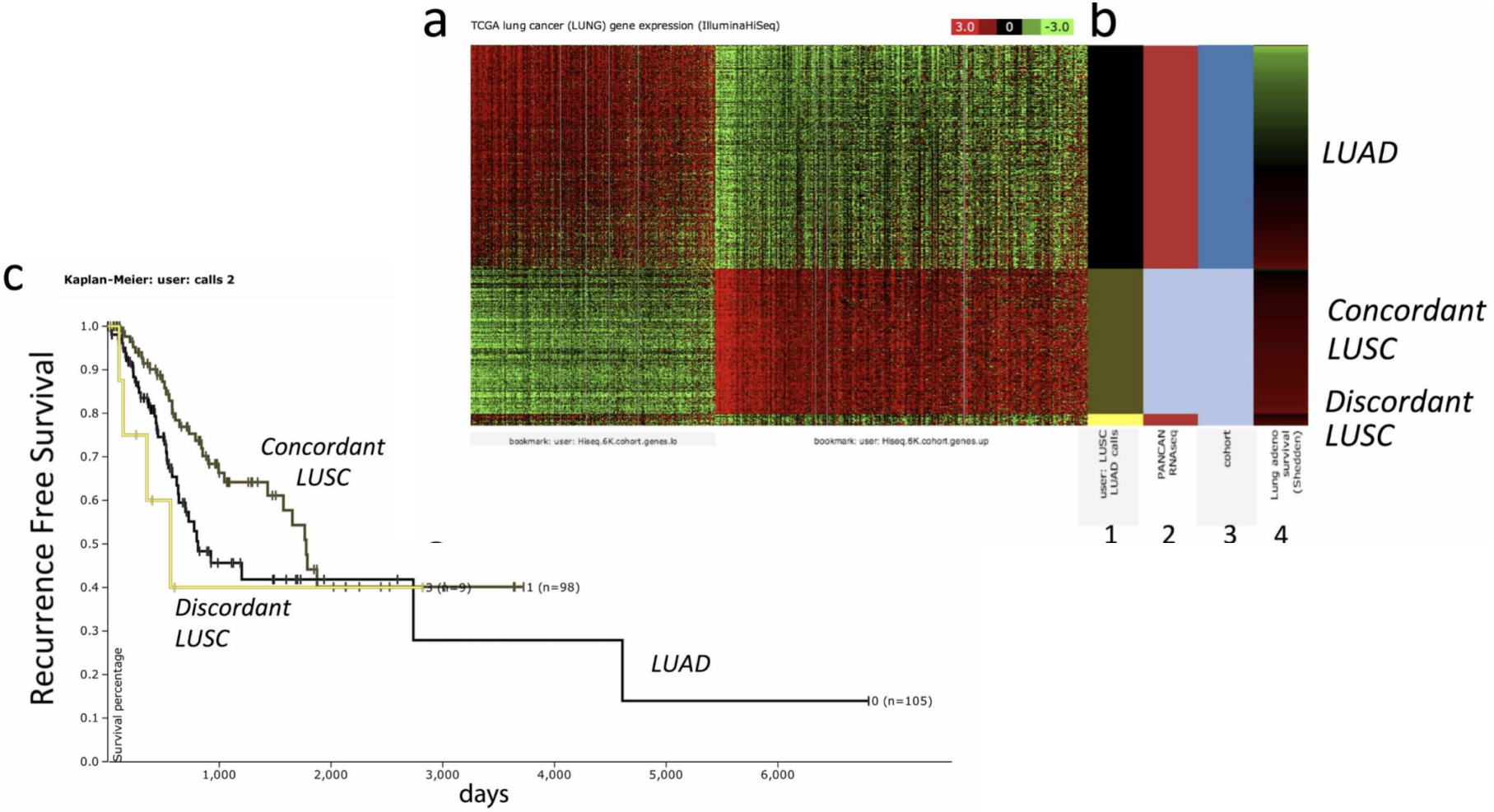
Exploring TCGA Pan-Cancer Data at the UCSC Cancer Genomics Browser

- Interactive visualization
- Exploration of TCGA genomic, phenotypic, and clinical data.

Researchers can explore the **impact of genomic alterations on phenotypes** by **visualizing**

- **Gene**
- **protein expression**
- **copy number**
- **DNA methylation**
- **somatic mutation**
- **Pan-Cancer subtype classifications and genomic biomarkers**

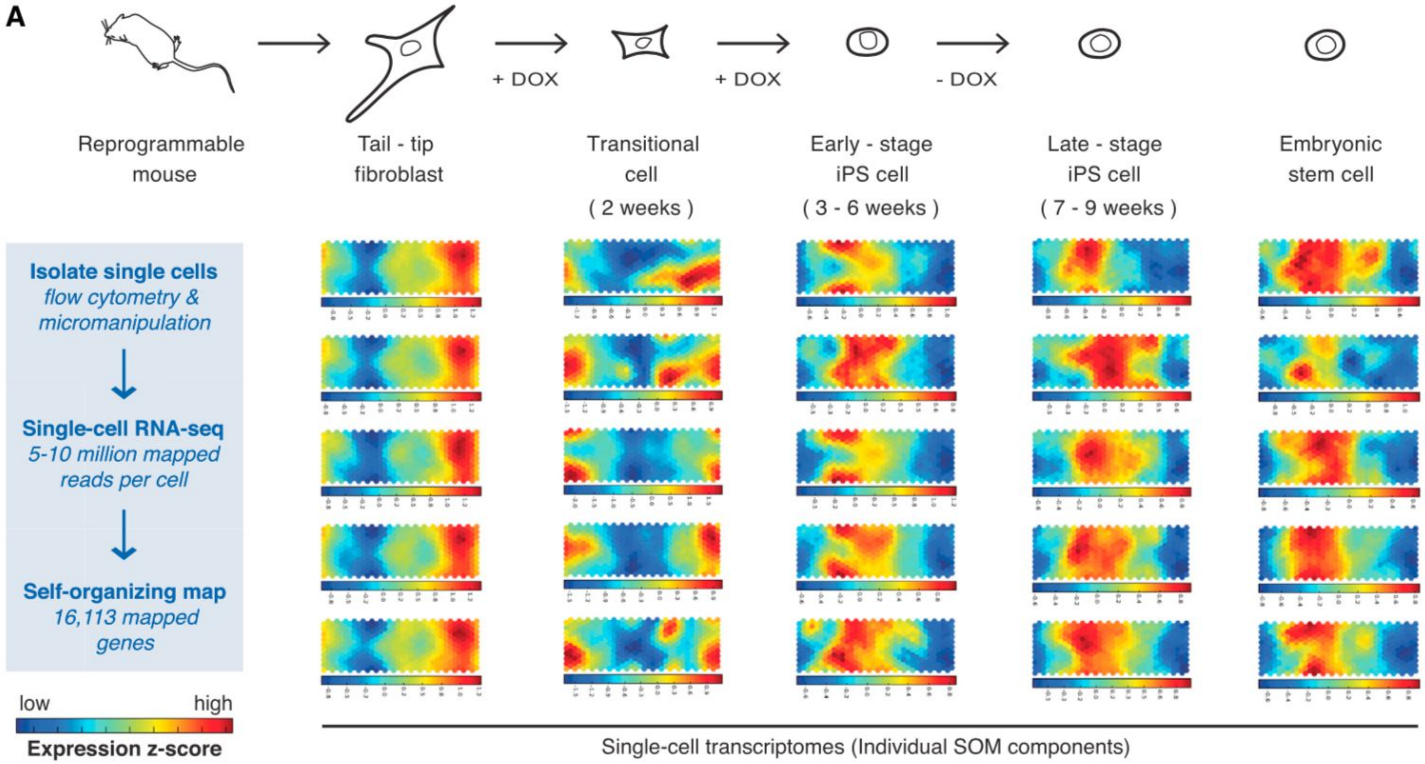

a
TCGA acute myeloid leukemia (LAML) somatic mutation • N=196
b
cluster3
cluster5
bookmark: user: LAML Significantly Mutated Genes
1
2
3
c
Kaplan-Meier: _LAML microRNA subtype (syn1688309)
Overall Survival
Survival percentage
Days
cluster5 (n=17)
cluster1 (n=14)
cluster2 (n=39)
cluster4 (n=55)
cluster3 (n=48)


a
TCGA lung cancer (LUNG) gene expression (IlluminaHiSeq)
3.0
0
-3.0
bookmark: user: Hiseq.6K.cohort.genes.lo
bookmark: user: Hiseq.6K.cohort.genes.up
b
LUAD
Concordant LUSC
Discordant LUSC
user: LUSC LUAD calls
PANCAN RNAseq
cohort
Lung adeno survival (Shedden)
1
2
3
4
c
Kaplan-Meier: user: calls 2
Recurrence Free Survival
Survival percentage
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Concordant LUSC
Discordant LUSC
LUAD
3 (n=9)
1 (n=98)
0 (n=105)
1,000
2,000
3,000
4,000
5,000
6,000
days

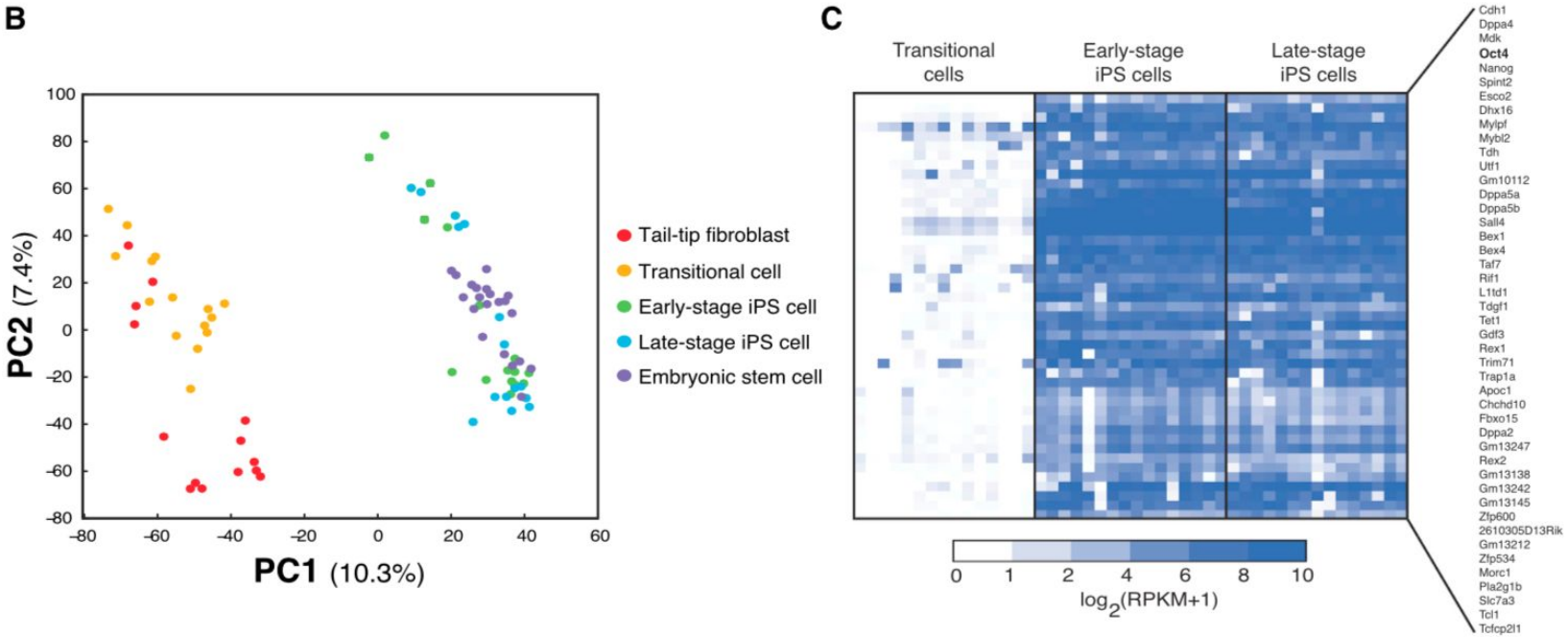
Single-Cell Transcriptome Analysis Reveals Dynamic Changes in lncRNA Expression during Reprogramming (2015)

- Looks at single cell transcriptomes during somatic cell reprogramming and characterizes long noncoding RNAs (lncRNA) at different stages of reprogramming.
- Analyzed their high-dimensional data using a self-organizing map (SOM).
    - Useful to visualize single cells based on the behavior of expressed gene sets

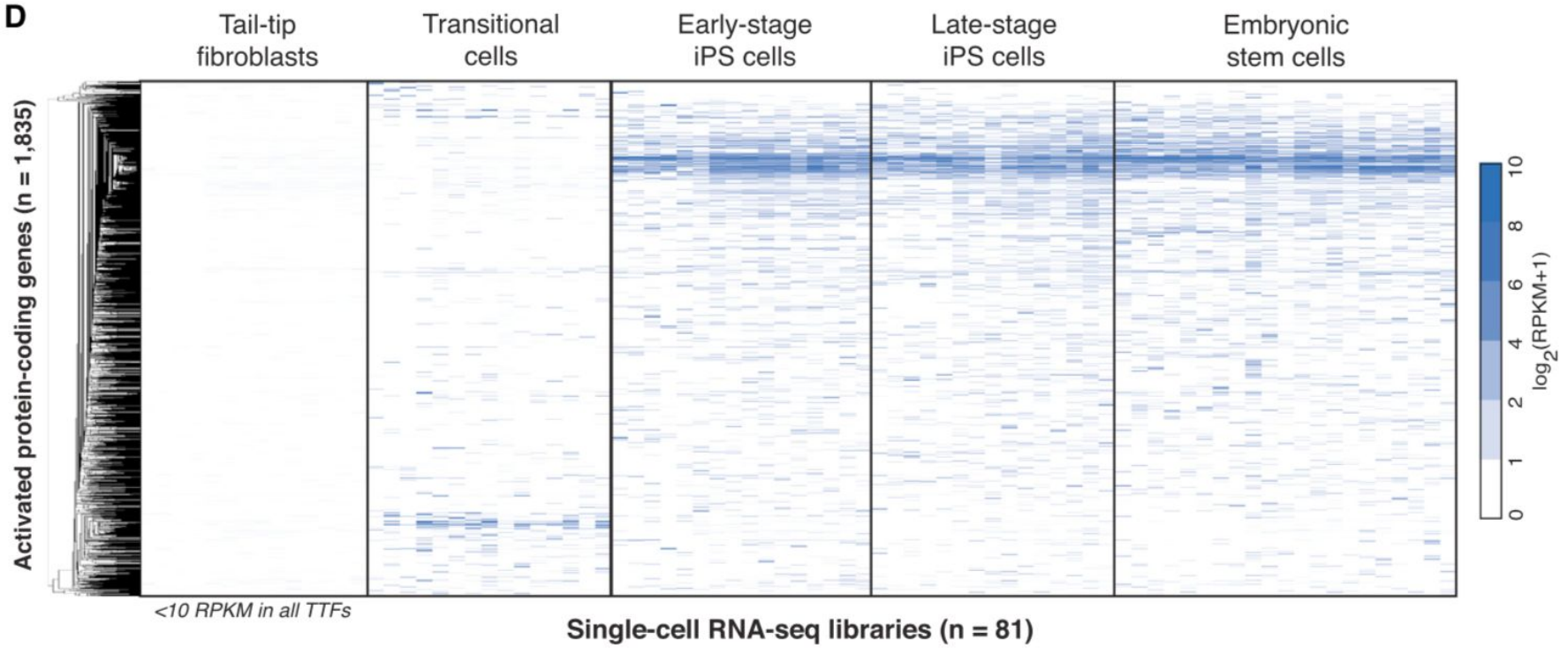
Single-Cell Transcriptome Analysis Reveals Dynamic Changes in lncRNA Expression during Reprogramming (2015)



Single-Cell Transcriptome Analysis Reveals Dynamic Changes in lncRNA Expression during Reprogramming (2015)



Single-Cell Transcriptome Analysis Reveals Dynamic Changes in lncRNA Expression during Reprogramming (2015)



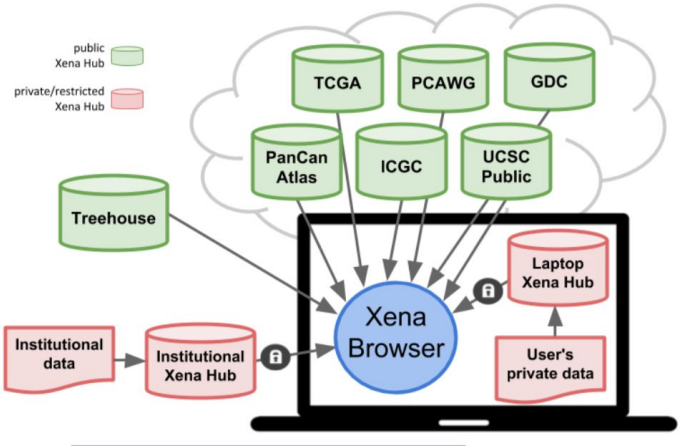
Single-Cell Transcriptome Analysis Reveals Dynamic Changes in lncRNA Expression during Reprogramming (2015)

Things noted:

- These visualizations are static, used primarily for publications rather than for data exploration
- There are many different types of separate visualizations used, rather than one single tool
- Aren’t dynamic--they don’t change when new data is added

The UCSC Xena Platform for cancer genomics data visualization and interpretation (2018)

- Web based visualization tool for genomic data, and clinical and phenotypic annotations
- Consists of the Xena browser in the front end and Xena hubs in the backend.
  - Xena hubs contain either public datasets from places such as The Cancer Genome Atlas (TCGA), Genomic Data Commons (GDC) and International Cancer Genome Consortium (ICGC), or private data sets supplied by the user
-



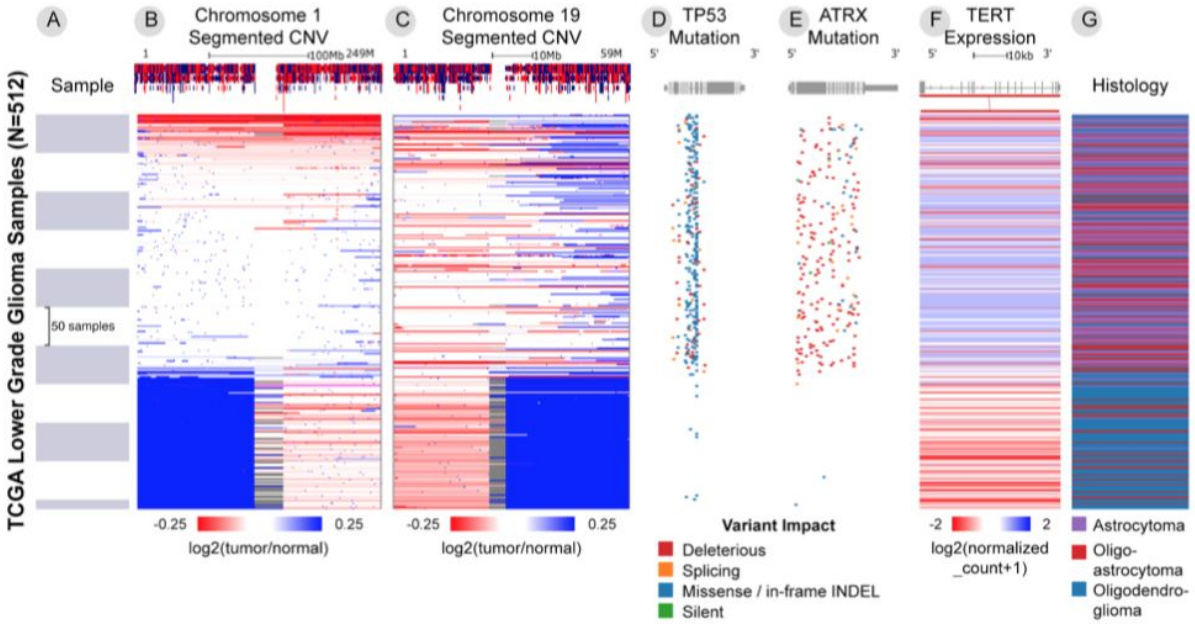
The UCSC Xena Platform for cancer genomics data visualization and interpretation (2018)

Motivation:

- Growing amount of genomics data, including different types of data
  - Somatic mutations, copy number, gene expression
- Data is highly distributed and used by researchers all over the world
  - Difficult to share data and connect various datasets
- Facilitate the comparison of different data to get better understanding of a genomic event

The UCSC Xena Platform for cancer genomics data visualization and interpretation (2018)

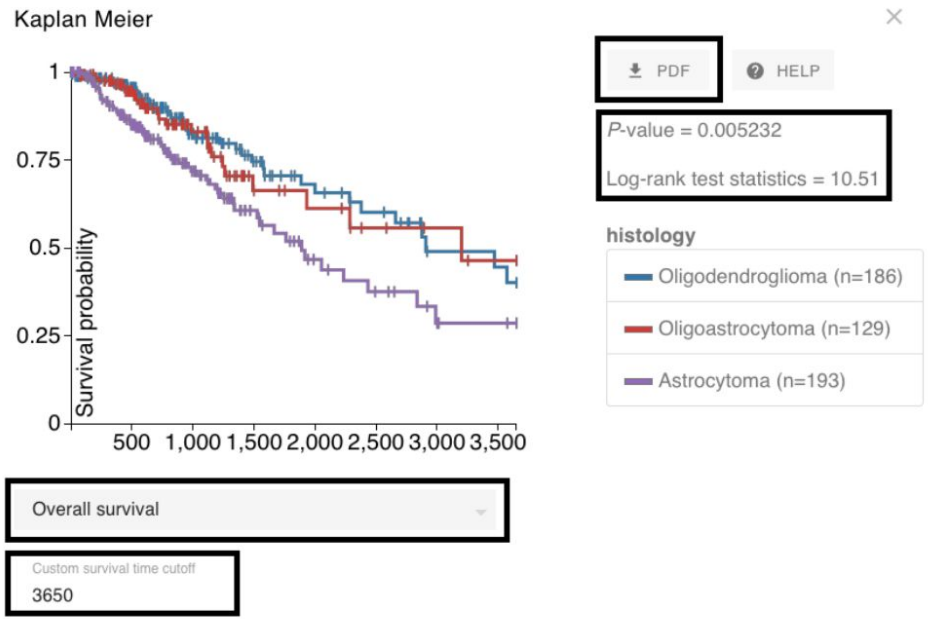
Xena Visual Spreadsheet



- Can view data side by side
- Each feature is a grid where each column is a slice of genomic or phenotypic data
  - E.g: gene expression or age
- Each row is a single entity
  - E.g: tumor sample

The UCSC Xena Platform for cancer genomics data visualization and interpretation (2018)

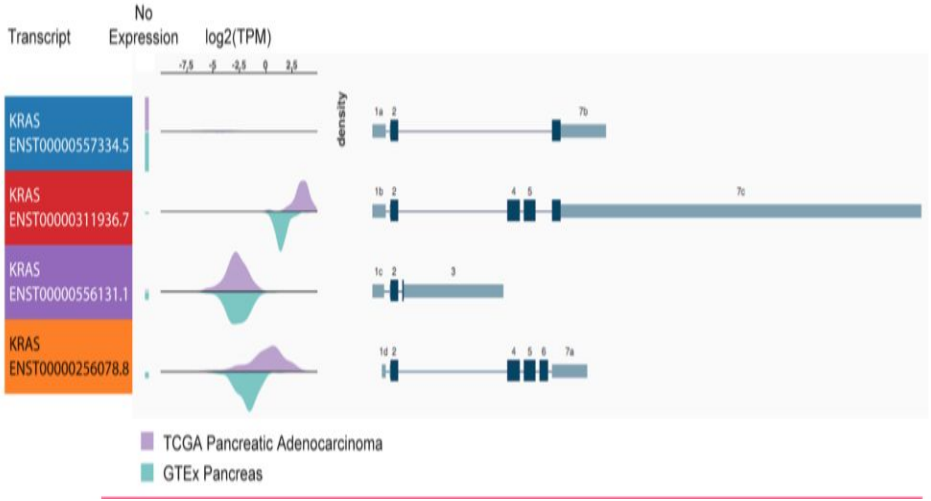
Kaplan-Meier analysis



- Visualization showing survival data for different histological subtypes, in this example there are 3
- Button to generate PDF
- Statistical analyses results
- Dropdown menu to select different survival endpoints
- Textbox to enter custom survival time cutoff

The UCSC Xena Platform for cancer genomics data visualization and interpretation (2018)

Transcript view



Enables users to compare transcript level expressions between two different groups of samples for all transcripts of genes in that sample

The UCSC Xena Platform for cancer genomics data visualization and interpretation (2018)

Other features:

- Ability to bookmark a current view to share with others via a generated link
- Box plots, bar charts, scatter plots
- Automatically computes chi-squared, t-test, and ANOVA
- Visualizations are integrated with the UCSC Genome Browser and other external sources
- Text based search to highlight, filter and group samples

A Taxonomy of Visualization Tasks for the Analysis of Biological Pathway Data (2016)

- Biological pathway data is used to represent chains of interactions
  - E.g: kRas and how it drives mutations in other cells, resulting in a cascade
- This paper develops a taxonomy based on interviews with biologists, aimed to support the design and development of biological pathway visualizations

A Taxonomy of Visualization Tasks for the Analysis of Biological Pathway Data (2016)

Table 2 A summary of the biological pathway visualization task taxonomy

Category	Example task
Attribute tasks	
(A1) Multivariate	Find all up-regulated genes in a biological pathway. Integrate results of a laboratory experiment into existing protein-protein interaction networks.
(A2) Comparison	Compare a biological pathway to a pathway with the same functionality in a reference species.
(A3) Provenance	Determine which studies provides the evidence for a link between two genes.
(A4) Uncertainty	Understand which pathway components have the strongest empirical evidence relationships.
Relationship tasks	
(R1) Attributes	Find all translocations of entities in a given biological pathway.
(R2) Direction	Find the products or output of a biochemical reaction.
(R3) Grouping	Expand a module entity to include all child-entities in the visualization.
(R4) Causality	Find all genes downstream of the currently selected entity, which may be affected by a change in regulation.
(R5) Feedback	Identify potential feedback loops in gene regulation.
Modification tasks	
(M1) Annotate	Update out-of date-information in a pathway data set, or create a personalized pathway relevant to a specialized research topic.
(M2) Curate	Identify errors and update historical data.

A Taxonomy of Visualization Tasks for the Analysis of Biological Pathway Data (2016)

Challenges:

- Genes, proteins and other molecules in a cell have very complex relationships with each other, including feedback loops. These can be hard to visualize effectively.
- Static and non-interactive visualizations often fail to convey dynamic information about a pathway
- Data is stored in many different formats that initially may not be set up in a way that easily connects them
- Complexity and large amounts of data that need to be included in the visualization make static images cluttered and hard to read

Transcriptomic Characterization of SF3B1 Mutation Reveals its Pleiotropic Effects in Chronic Lymphocytic Leukemia

- SF3B1 gene mutation is known to affect the progression of CLL cells in humans
- This paper identifies specific pathways affected by the mutation to isolate why they affect the progression of these leukemia cells
- Published in Cancer Cell Journal, 2016
- Authored by many researchers from various institutions including Angela Brooks (UCSC)

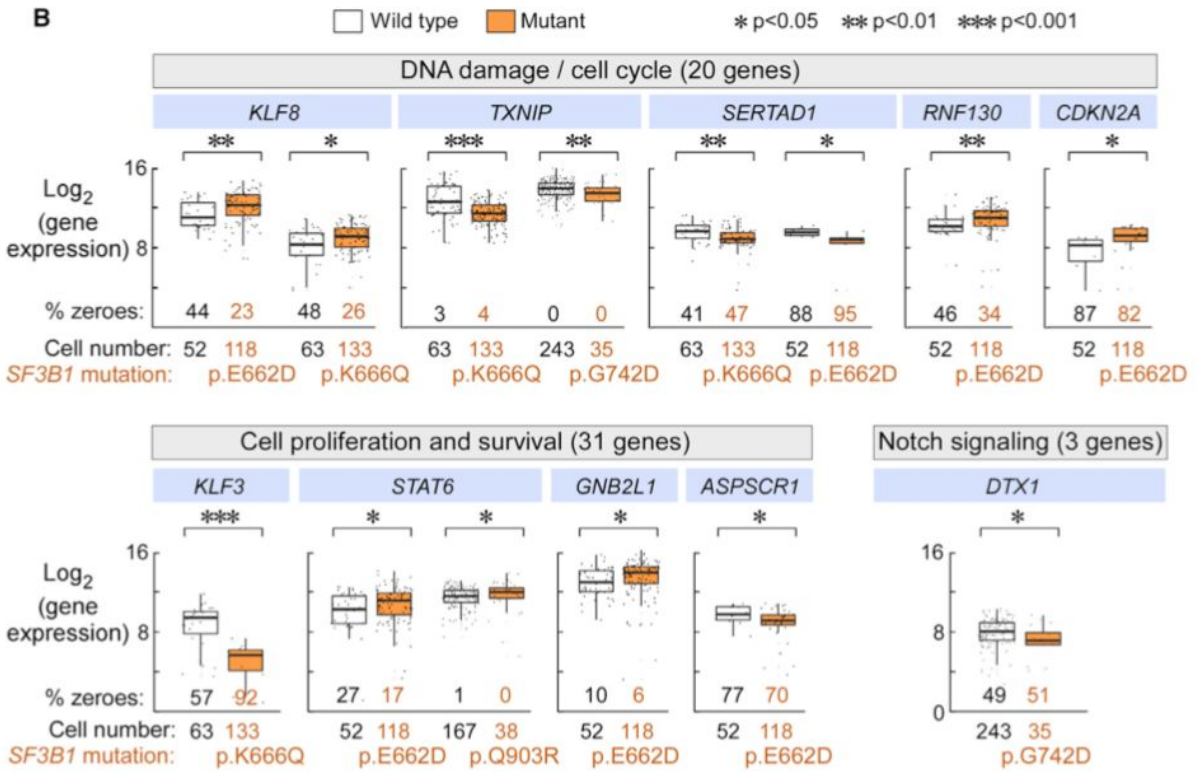
Transcriptomic Characterization of SF3B1 Mutation Reveals its Pleiotropic Effects in Chronic Lymphocytic Leukemia

Experiment

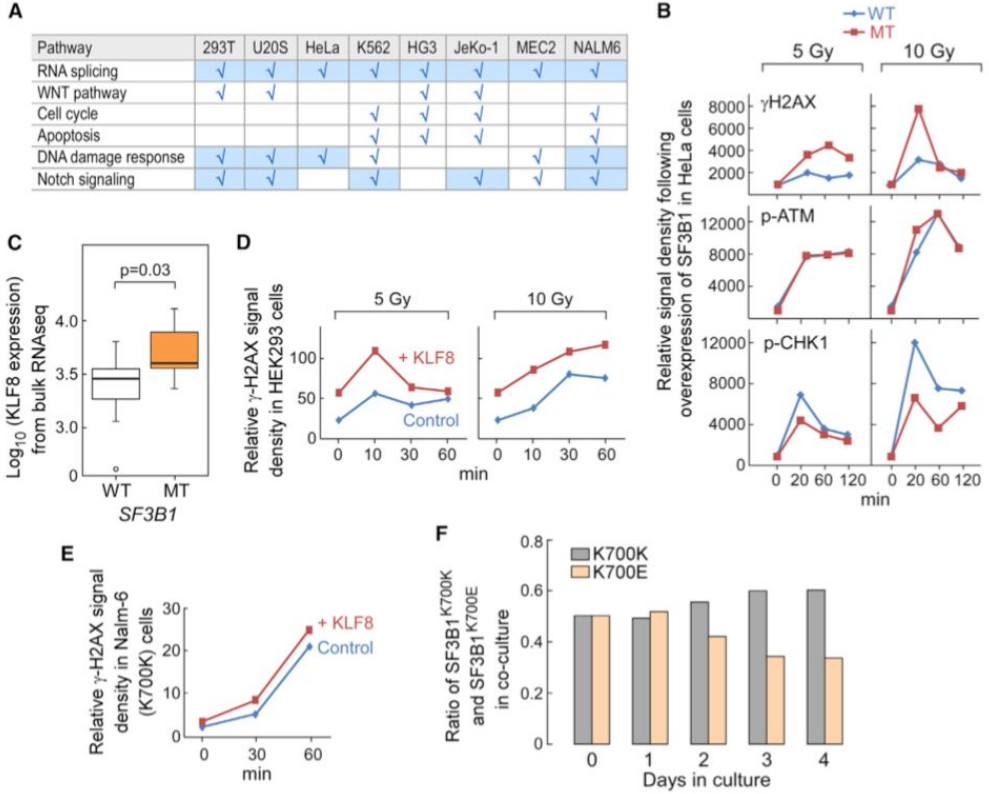
- Blood samples were taken from healthy donors and used to identify mutation status of SF3B1
- Transcriptomic Characterization - studying the complete set of RNA transcripts that are produced by the genome

- The mutations in SF3B1 in normal stem cells is not the main cause of cancer in cells
- Mutation in SF3B1 alongside other cancer-driving alterations accelerates CLL cell development

Transcriptomic Characterization of SF3B1 Mutation Reveals its Pleiotropic Effects in Chronic Lymphocytic Leukemia



Transcriptomic Characterization of SF3B1 Mutation Reveals its Pleiotropic Effects in Chronic Lymphocytic Leukemia



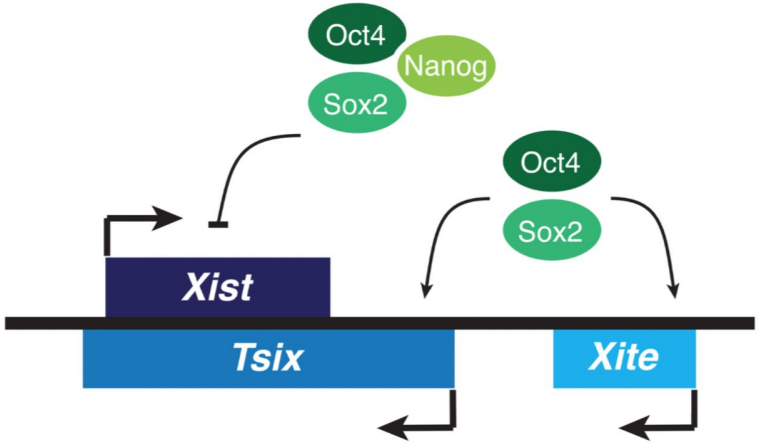
X-chromosome epigenetic reprogramming in pluripotent stem cells via noncoding genes

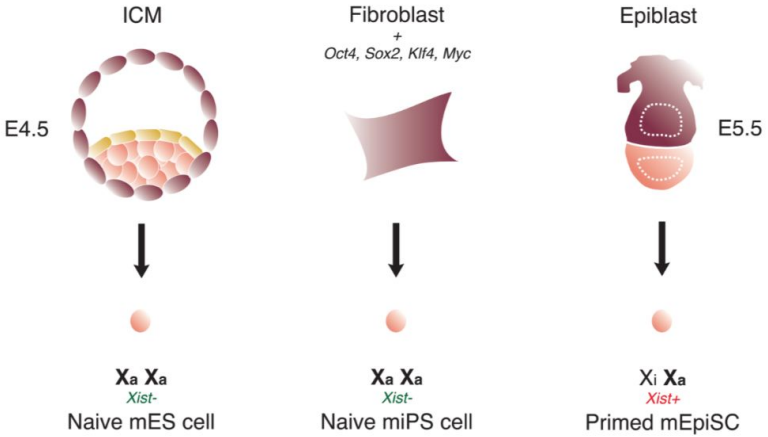
- Discusses the potential effect of x-chromosome status to provide benchmarks on the epigenetic quality of pluripotent stem cells
  - Epigenetic - relative to or arising from nongenetic influences on gene expression
  - Pluripotent - not yet developed into mature cells
- Authored by many including Daniel Kim
- Published in Seminars in Cell and Developmental Biology, 2011

X-chromosome epigenetic reprogramming in pluripotent stem cells via noncoding genes

- Experiments on mouse cells indicates evidence that there is an affinity between X-chromosome status and pluripotent stem cells
- “Deciphering the molecular mechanisms underlying X-chromosome reprogramming may yield new insights into the acquisition of pluripotent ground state”
- The reasoning behind why these changes occur is not well understood, and is proposed as a topic of further research in stem cell therapy and regenerative medicine

X-chromosome epigenetic reprogramming in pluripotent stem cells via noncoding genes

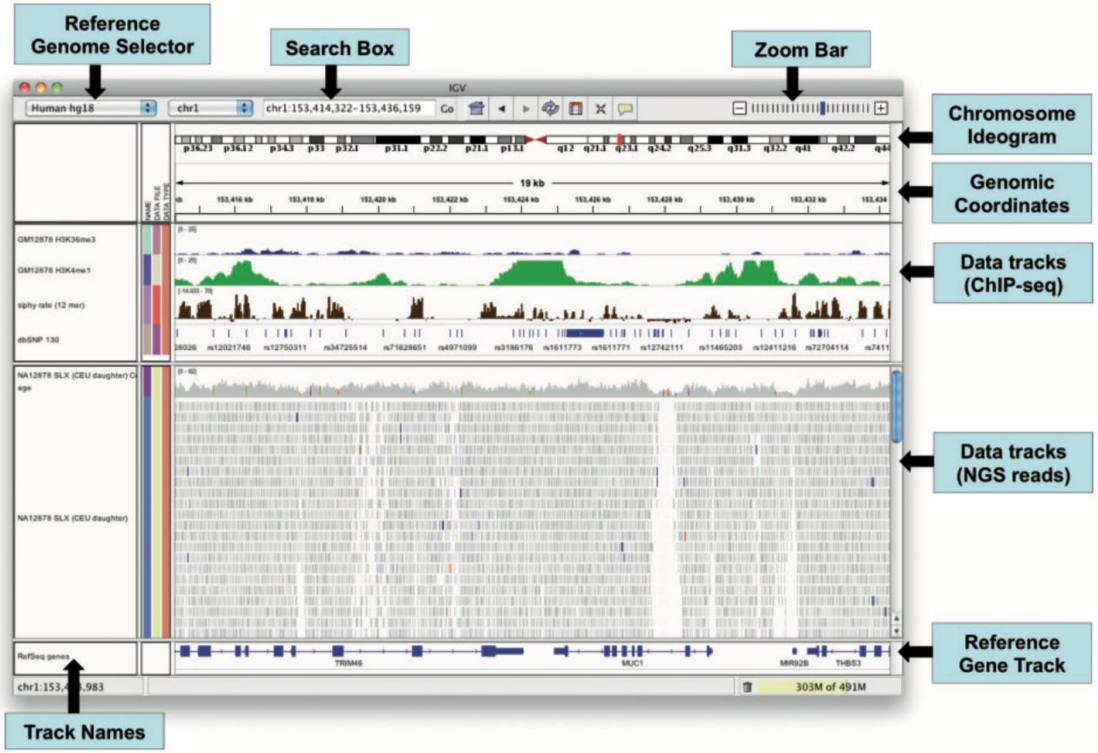




Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration

- Free software for researchers in genomics to visualize their own datasets with richer tools and approaches than previously available to match the complexity of today's sequencing methods
- Developed at the Broad Institute of MIT and Harvard
- Helga Thorvaldsdottir, James T. Robinson, and Jill P. Mesirov
- Bioinformatics, 2012

Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration



Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration

- The IGV allows users to view their own genomic datasets with a high level of flexibility in genome resolution
- Desktop application made for researchers that utilizes some publicly available datasets, but can also visualize any custom data that can be mapped to genomic coordinates
- Future plans (at time of publication)
  - Data driven search and navigation ability
  - Network visualizations for pathways
  - New approaches for understanding overall trends while still allowing access to lower level details