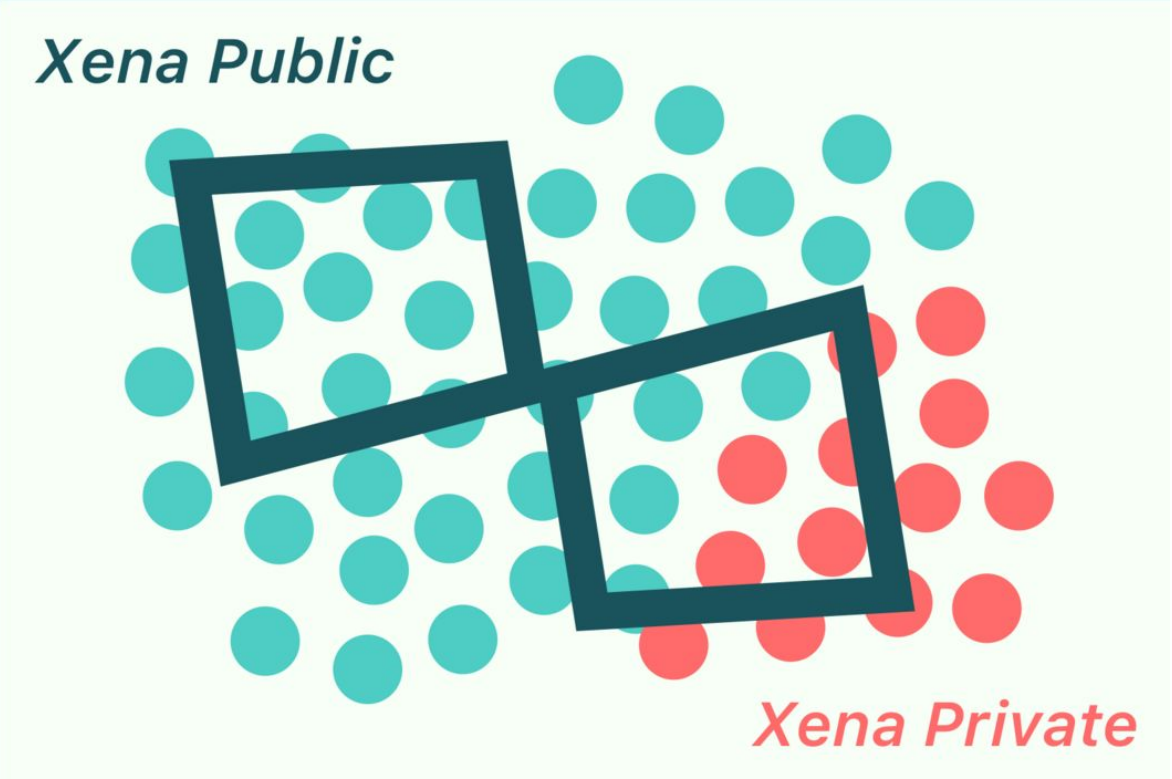
UCSC Xena

A Platform for Functional Genomics Visualization and Analysis



Mary Goldman
University of California Santa Cruz Genomics Institute
UCSC Xena Workshop for COH, April 19th, 2022

Agenda

1:30p - 2:30p Workshop Presentation and Demo

2:30p - 3:00p Q&A, break, set up for Hands-on section

3:00p - 5:00p Hands-on section

What is Xena?

Xena is a cancer functional genomics visualization and analysis tool for cohorts of samples

Xena is a **cancer functional genomics** visualization and analysis tool for cohorts of samples

Xena is a **cancer functional genomics visualization and analysis tool** for cohorts of samples

Xena is a **cancer functional genomics visualization and analysis tool** for **cohorts of samples**

Xena's Strengths

- Easy access to public cancer genomics resources and can view your own data or data from the literature
- Visual multi-omics data integration
- Many visualizations and analyses: Visual Spreadsheet, KM plot, dynamically make subgroups, differential expression analysis, violin plot, box plot, statistics, gene expression signatures

Questions you can ask with Xena ...

- Is overexpression of this gene associated with lower/higher survival?
- Is this gene differentially expressed in tumor vs normal samples?
- What are the top 10 differentially expressed genes between my two subgroups?

Highlights of some of our Public Data Resources

Example data types

- SNPs and small INDELs
- Large structural variants
- Segmented copy number, gene-level copy number
- Gene expression; Transcript-, Exon-, Protein-, LncRNA-, and miRNA-expression
- DNA methylation (array and WGBS)
- ATAC-seq peak signal
- Phenotype, clinical data, survival endpoints
- Signature scores, classifications

TCGA

- 40 cancer types, 12,000 samples
- Many types of data: somatic mutation, gene expression, copy number, and more
- Survival + other basic phenotype/clinical data

4 versions of TCGA data in Xena:

1. Newest data from the PanCan Atlas project
2. Harmonized data from the GDC
3. Legacy data published when the TCGA data originally came out
4. UCSC RNAseq compendium

TCGA + TARGET (pediatric cancer) + GTEx (normal tissues) all from the same computational pipeline. Some people use it compare tumor vs normal.

Genomic Data Commons (GDC)

- Variety of harmonized datasets for a growing number of projects
- Xena has TCGA, TARGET, MMRF-COMPASS
  - TARGET: RNAseq and copy number data for pediatric samples
  - MMRF-COMPASS: longitudinal data from multiple myeloma patients
- CPTAC and others coming soon!

UCSC RNA-seq Compendium (Toil)

Uniformly analyzed

TCGA + TARGET + GTEx +

KidsFirst RNA-seq data

Published: 11 April 2017

Toil enables reproducible, open source, big biomedical data analyses

John Vivian, Arjun Arkal Rao, [...]Benedict Paten

Nature Biotechnology **35**, 314–316(2017) | Cite this article

(GTEx: expression for normal tissues)

Can be used to compare tumor vs normal

- 20K samples
- 60K genes, 200K transcripts expression

- **I**nternational **C**ancer **G**enome **C**onsortium
- TCGA data + more studies from around the world (15,000 samples)
- Non-TCGA samples have non-coding mutations

Xena Study: 'ICGC (donor centric)'

Pan-**C**ancer **A**nalysis of **W**hole **G**enomes

Article | Open Access | Published: 07 July 2020

A user guide for the online exploration and visualization of PCAWG data

Mary J. Goldman, Junjun Zhang, Nuno A. Fonseca, Isidro Cortés-Ciriano, Qian

- Whole-genome! (most data is exome only)
  - Very rich data source
  - Many different types of -omics data
- Samples selected from ICGC (2,000 samples)

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium Nature **578**, 82–93(2020)

- **C**ancer **C**ell **L**ine **E**ncyclopedia
- Genetic and pharmacologic characterization of 1100 cell lines
- CNV, Expression, Somatic mutation, drug response

Xena Study: 'Cancer Cell Line Encyclopedia (CCLE) '

MET500

- Robinson, et al. Integrative clinical genomics of metastatic cancer. Nature (2017)
- Metastatic Tumor Cohort
  - 500 samples

Xena Study: 'MET500 (expression centric)' study

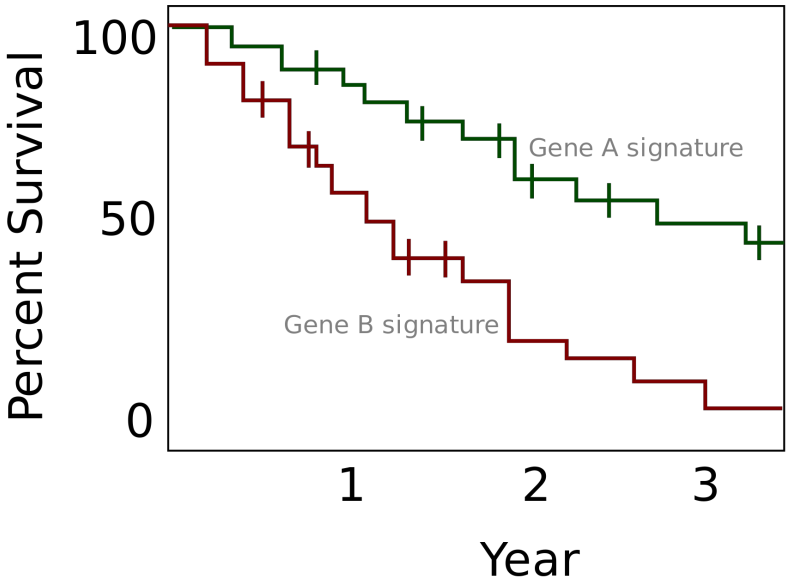
Overview of Visualizations

Visualizations

- Xena visual spreadsheet. Rows are samples and columns are genome-wide data
- See relationships between different types of data
- Allows you to dive into the tumor's biology

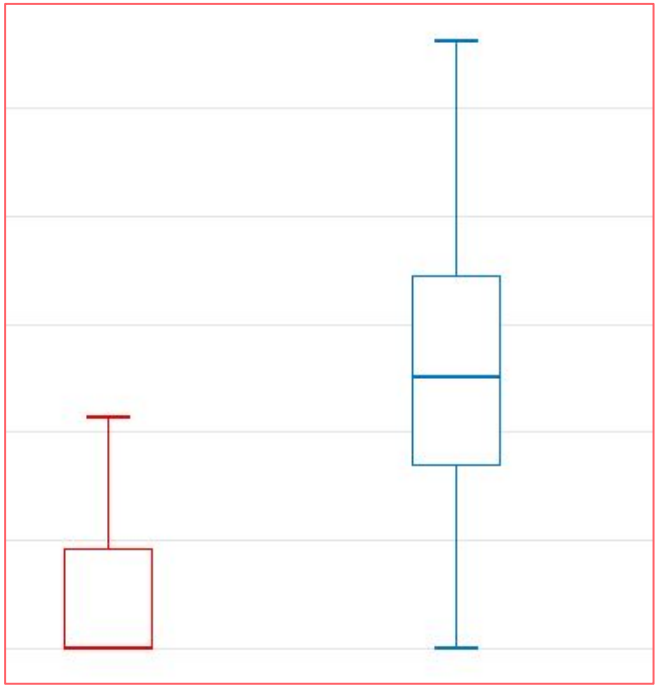
Visualizations

- KM plot with stats
- Analyze survival differences between groups of patients
- Steeper curve = worse prognosis



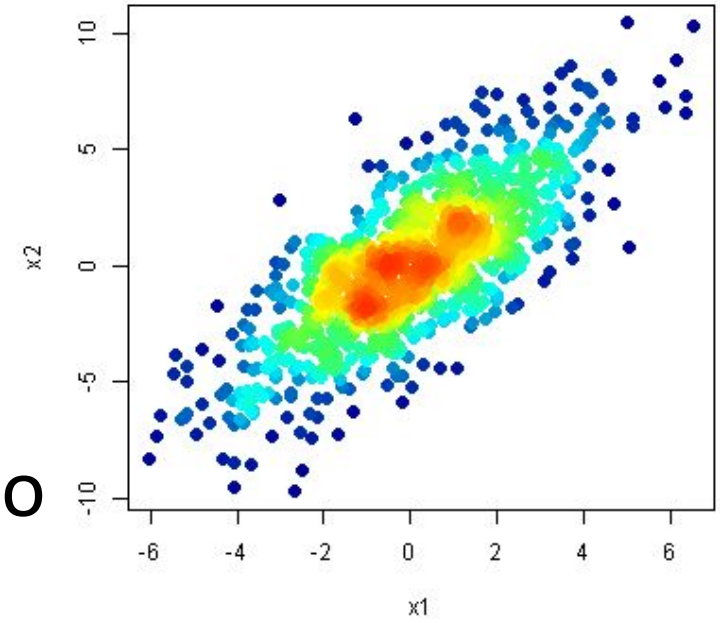
Visualizations

- Boxplots and violin plots with stats
- Compare groups of samples
- e.g. comparing average gene expression between tumor and normal



Visualizations

- Scatterplot with stats
- Great for seeing the relationship between two continuous variables
- e.g. compare expression to copy number variation



Shareable Live views

- Found something interesting and want to share?
- Make a bookmark URL and send to anyone
  - Drops them into a live view for further exploration
  - e.g. https://xenabrowser.net/heatmap/?bookmark=aaf8954f9b2b0577bc87a4334a1ca7bf
- Can also make a PDF for a presentation or publication

Example analysis: *Molecular subtypes in brain cancers*

2 Subgroups in Lower Grade Glioma:

- Characterized by loss of chromosome 1p and 19q
- Better survival prognosis

- Characterized by mutations in ATRX and TP53
- Worse survival prognosis

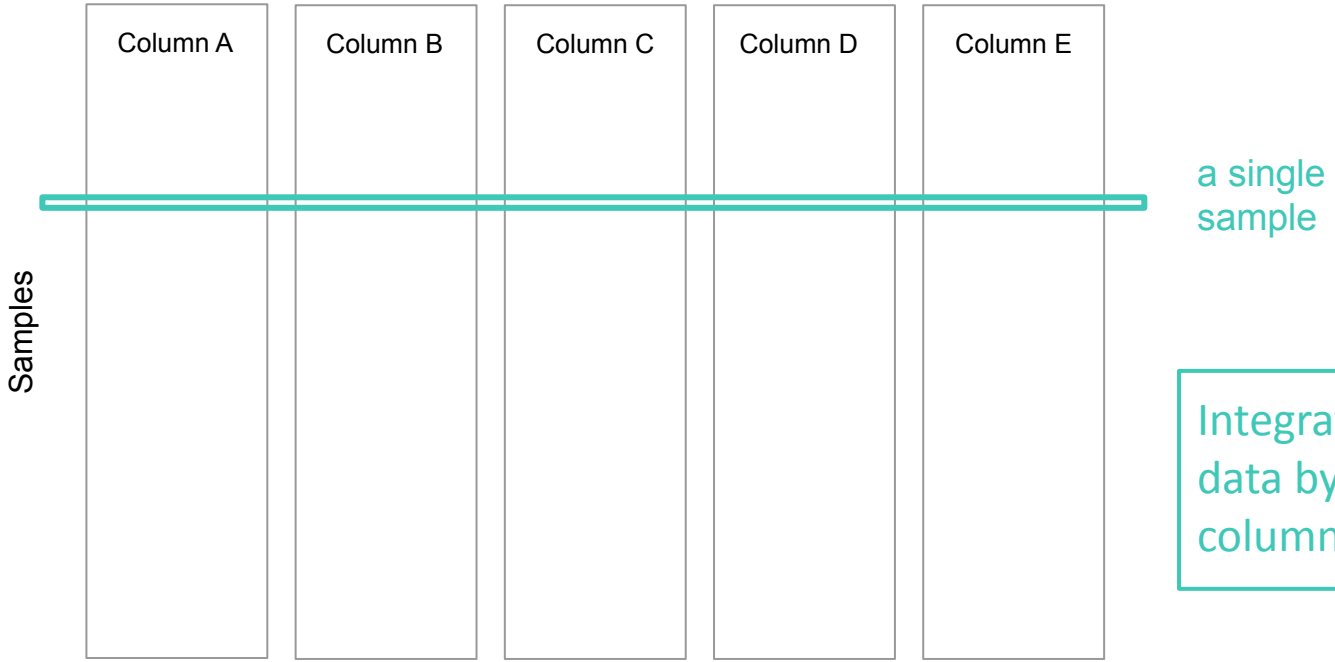
2 Subgroups in Lower Grade Glioma:

- Characterized by loss of chromosome 1p and 19q
- Better survival prognosis

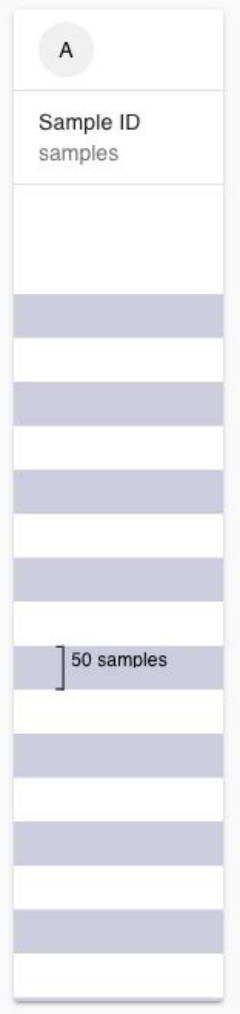
- Characterized by mutations in ATRX and TP53
- Worse survival prognosis

→ Do we see these subgroups in other brain cancers, like Glioblastoma Multiforme?

Xena Visual Spreadsheet



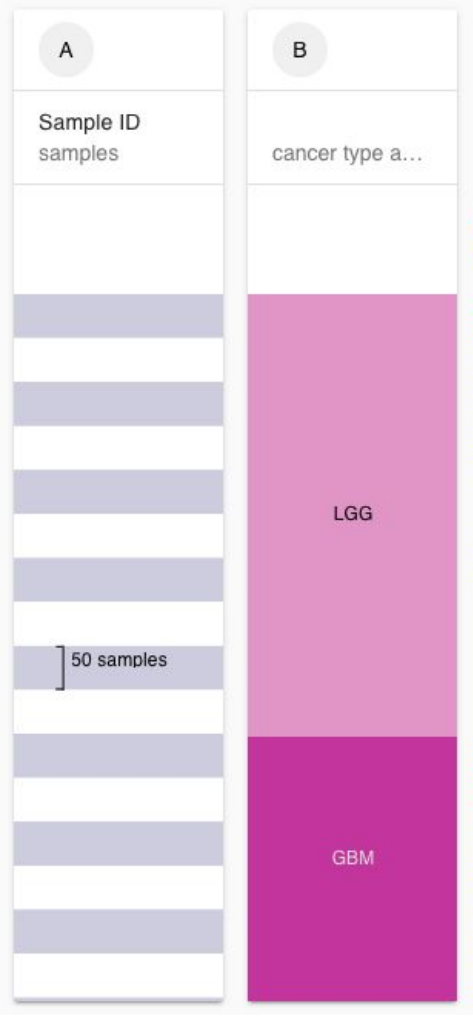
Integrate multi-omics data by placing data columns side by side



Samples

Each bar is 50 samples

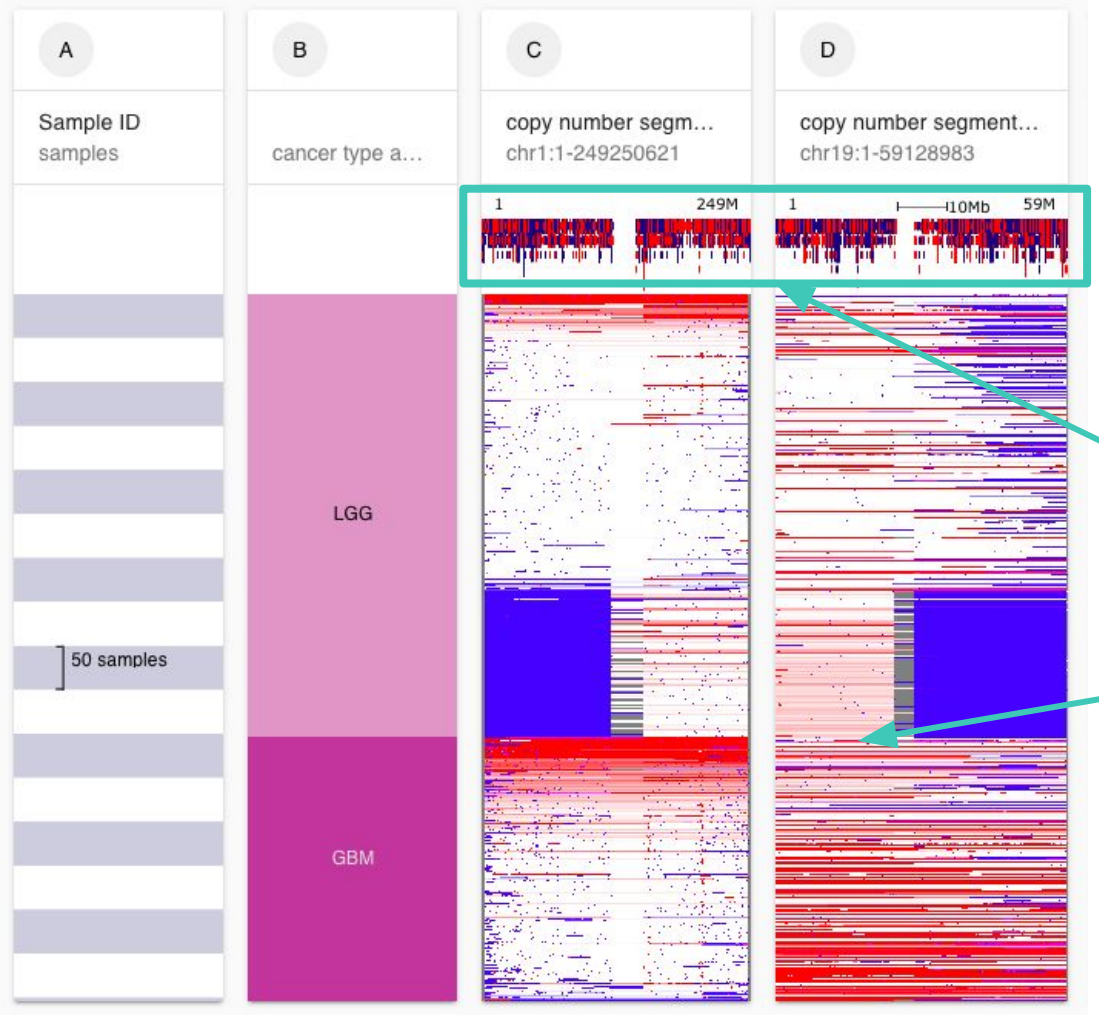
→ There are ~800 samples in view



Two types of cancer:

LGG = Lower Grade Glioma

GBM = Glioblastoma Multiforme



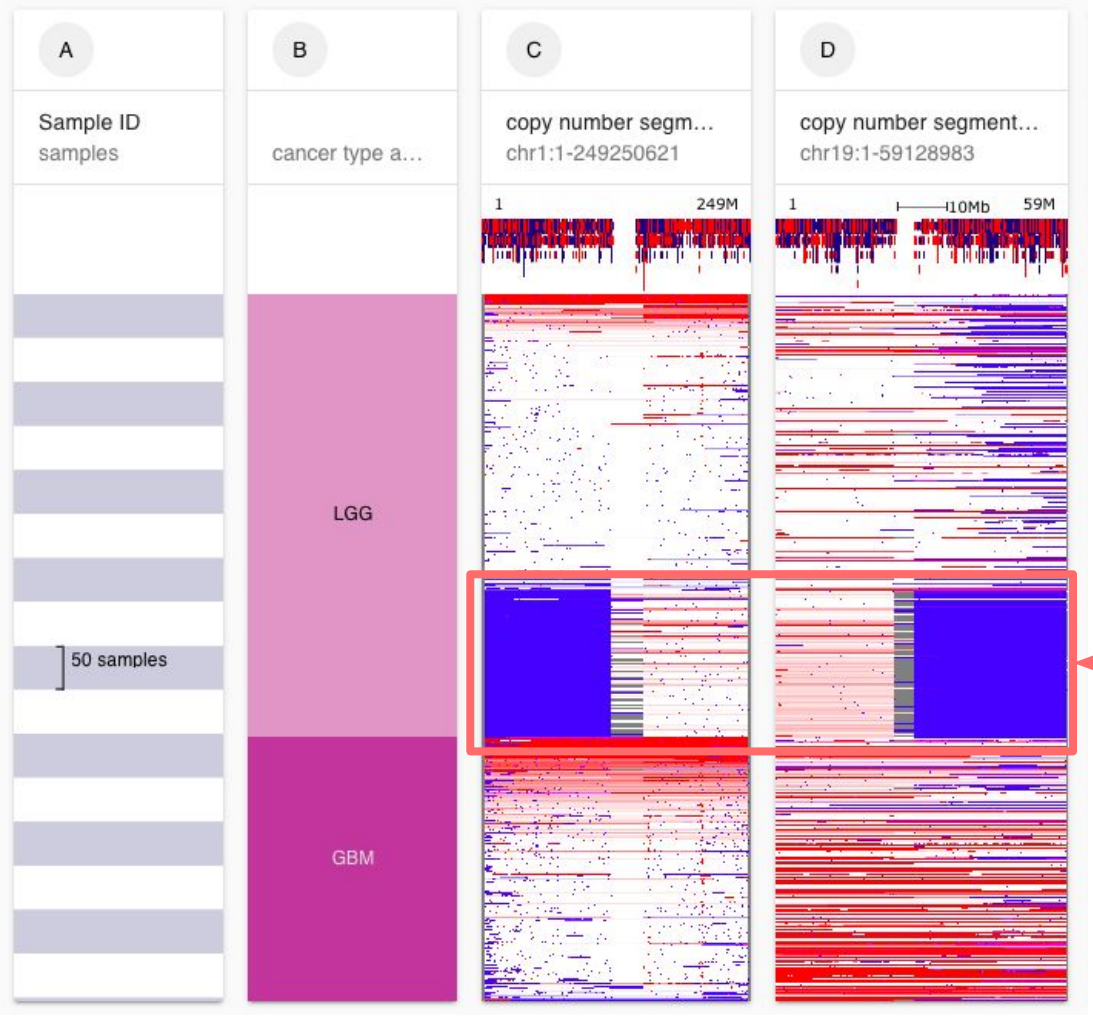
Copy Number Variation

Genes annotated at top in red (forward strand) and blue (reverses strand)

Red = Amplification

White = Neutral

Blue = Deletion



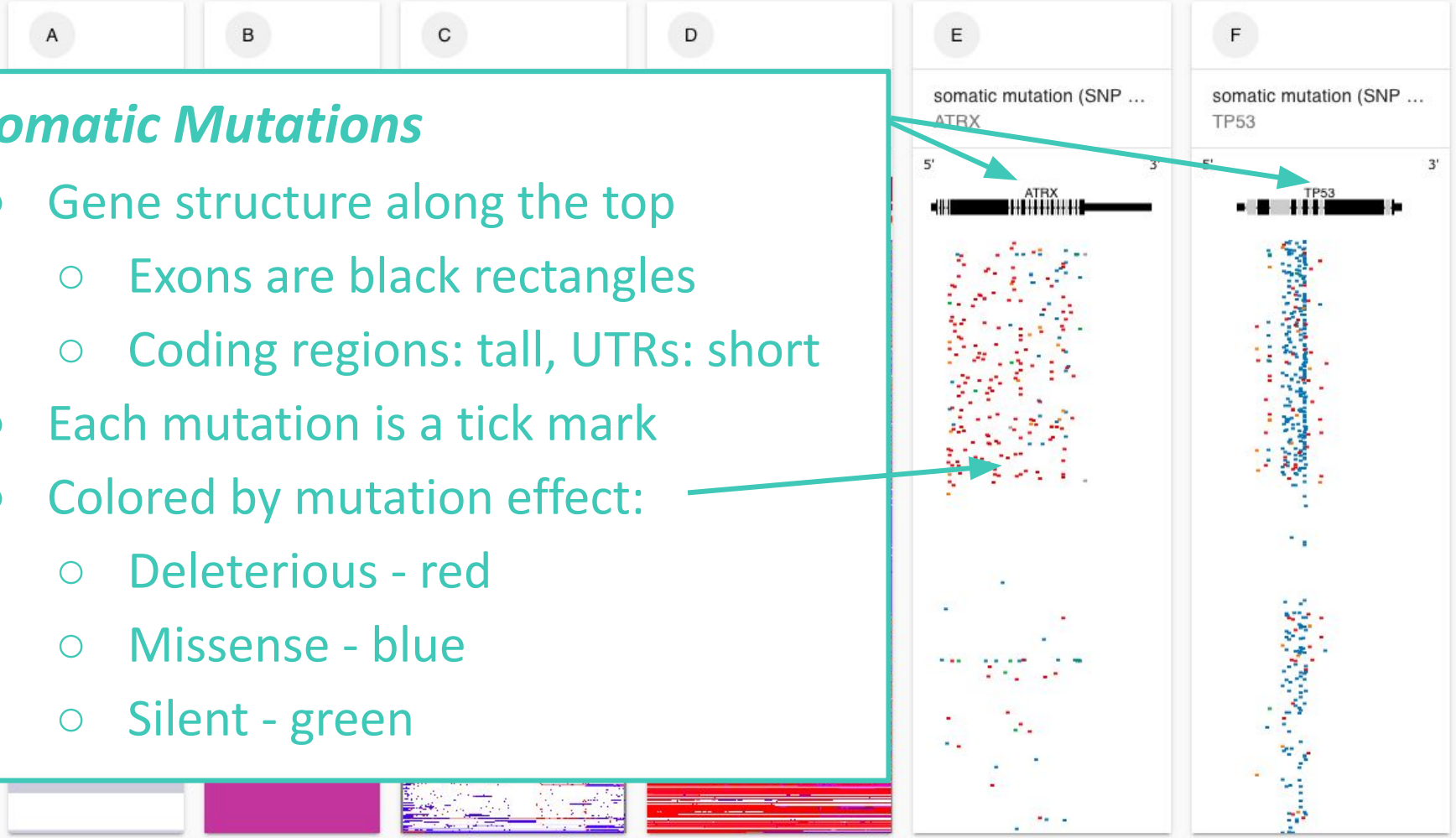
Copy Number Variation

Some LGG samples are characterized by co-deletion of chr1p and chr19q

We do not see this in GBM

Somatic Mutations

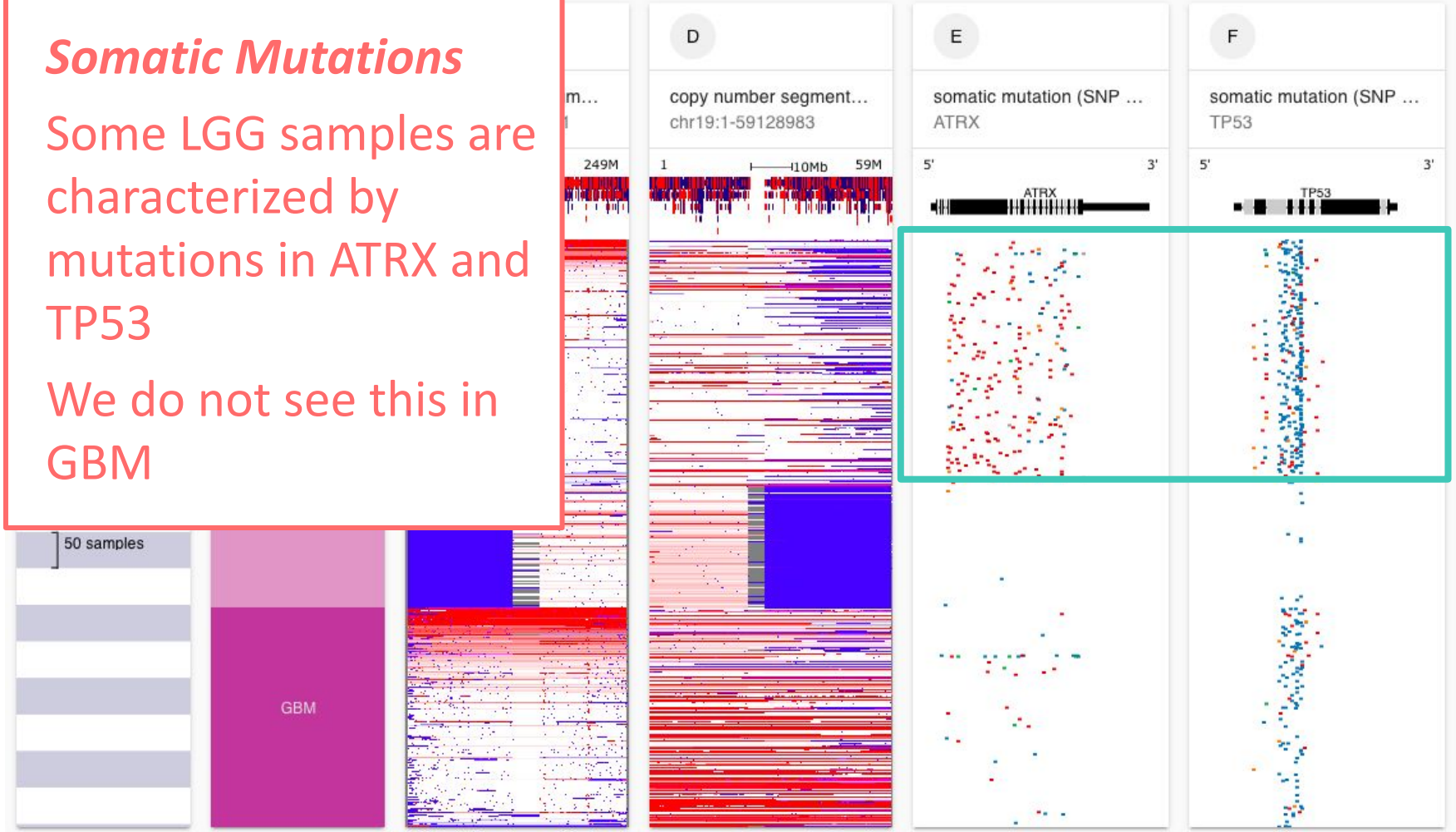
- Gene structure along the top
  - Exons are black rectangles
  - Coding regions: tall, UTRs: short
- Each mutation is a tick mark
- Colored by mutation effect:
  - Deleterious - red
  - Missense - blue
  - Silent - green



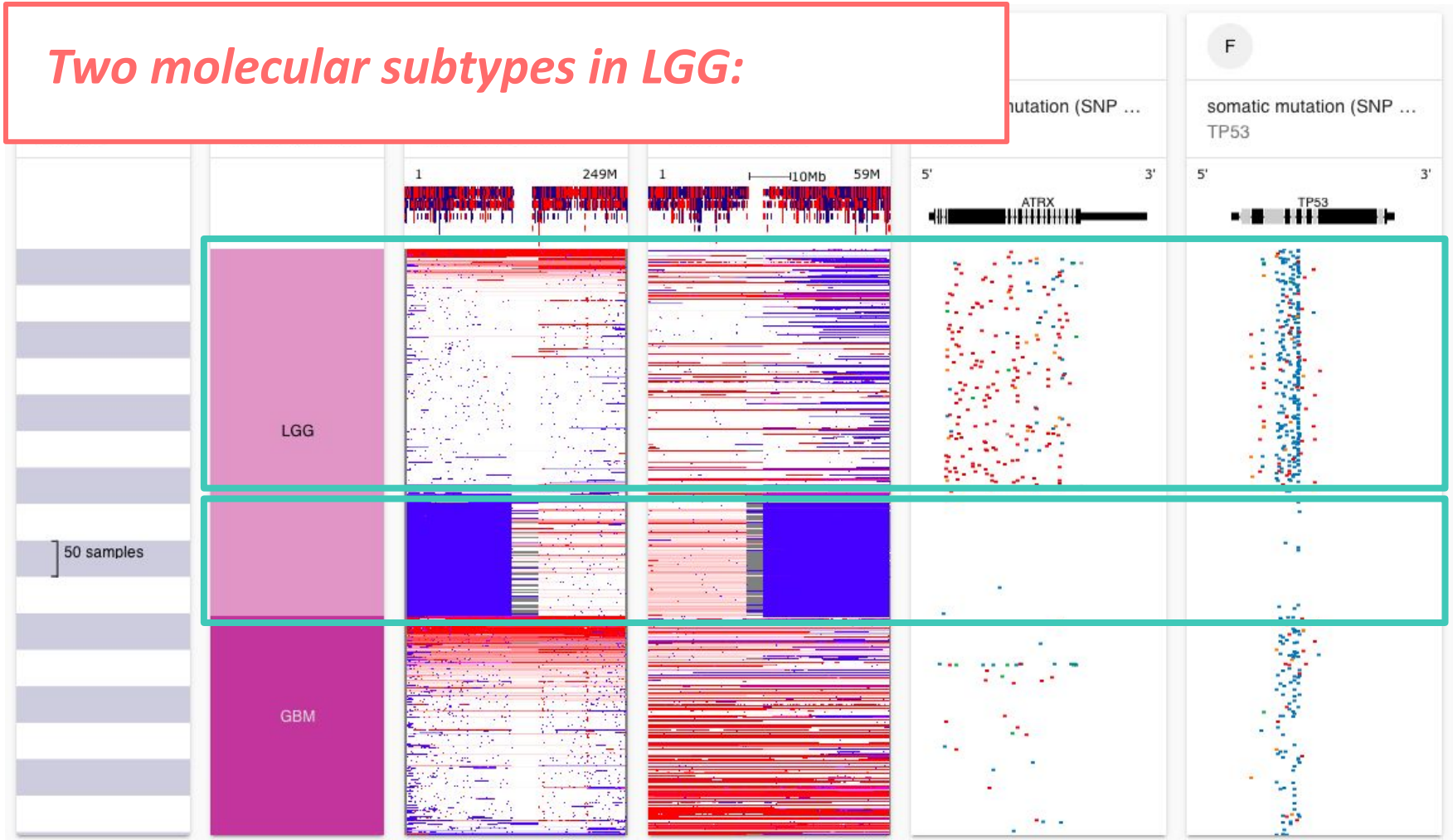
Somatic Mutations

Some LGG samples are characterized by mutations in ATRX and TP53

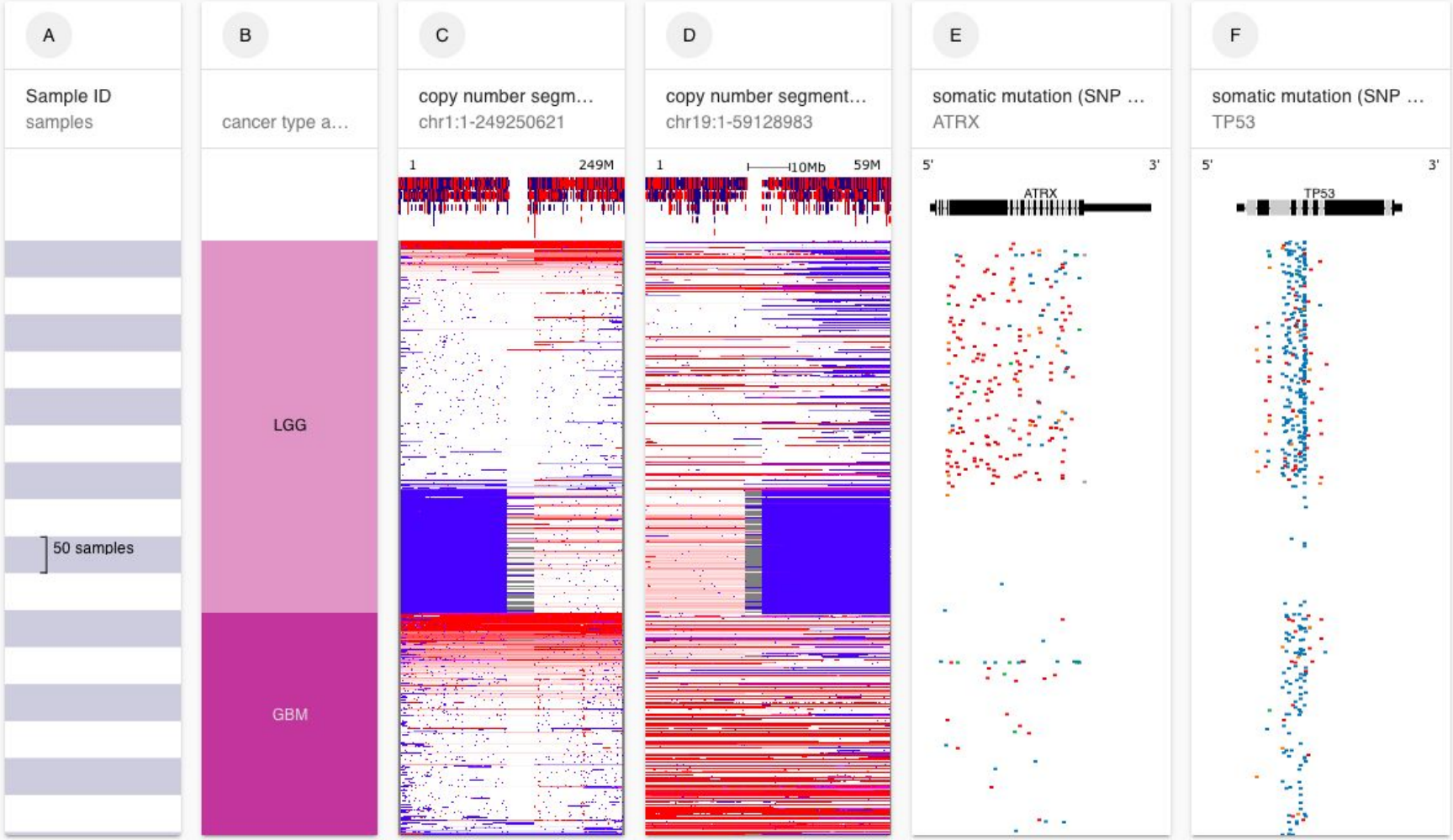
We do not see this in GBM



Two molecular subtypes in LGG:



We don't see these subtypes in GBM …

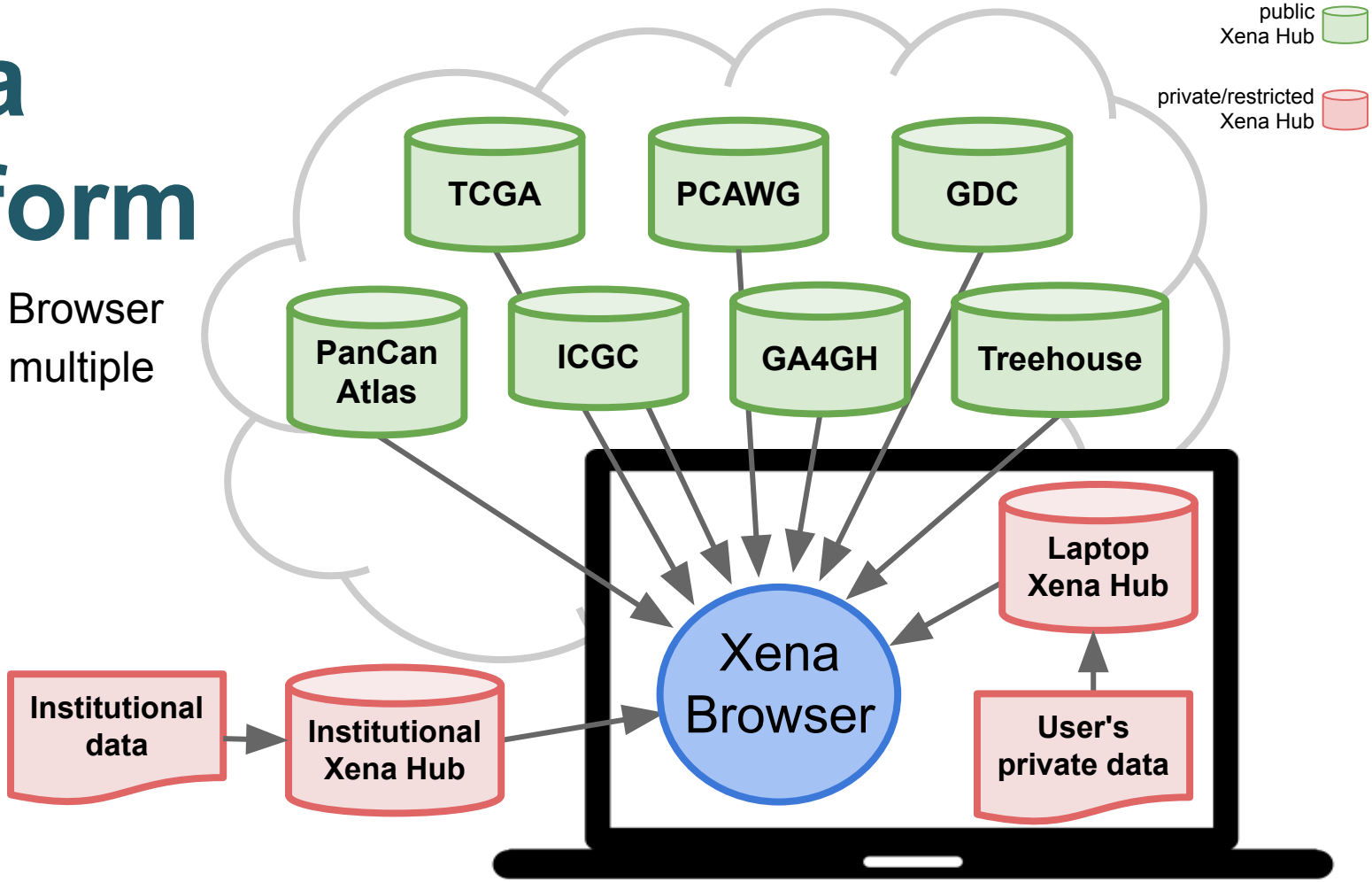

A
Sample ID
samples
50 samples
B
cancer type a…
LGG
GBM
C
copy number segm…
chr1:1-249250621
1
249M
D
copy number segment…
chr19:1-59128983
1
10Mb
59M
E
somatic mutation (SNP …
ATRX
5'
3'
ATRX
F
somatic mutation (SNP …
TP53
5'
3'
TP53

Demo

Viewing your own data

Xena Platform

Web-based Browser connects to multiple data hubs



Viewing your own data

- Can be:
  - new set of samples
  - more data on samples we already have on our public hubs
- We never upload your data to a public server - data stays on your computer
- You control data access. Can be for you only, collaborators only, or the public

Treehouse Public Xena Hub

Treehouse
CHILDHOOD CANCER INITIATIVE

The goal of the Treehouse Childhood Cancer Initiative (Treehouse) is to evaluate the utility of comparative gene expression analysis for difficult-to-treat pec
has now assembled a large collection of pediatric cancer RNA-Seq, which, added to adult data, results in a compendium of over 11,000 adult and pediatric
from public repository samples and from clinical samples at partner institutions, including UC San Francisco, Stanford, Children's Hospital of Orange Count
Genomics Institute's commitment to sharing data and to furthering research everywhere, we have made this data available for all to download and use.

We encourage the downloading and use of this data, and welcome any feedback you can offer to us to improve the user experience.

Treehouse is the pediatric cancer research arm of the UC Santa Cruz Genomics Institute. Under the leadership of Professor David Haussler, following the
distribute sequence data for TCGA, TARGET, and other large NCI genomics projects, UCSC Genomics Institute investigators extended this work by adding
adult cancer tumors. Treehouse and partner institutions are dedicated to exploring how computational methods can lead to better and safer treatment optio
expression data of each patient's tumor in the context of thousands of pediatric and adult tumors that have undergone similar characterization. Our genomi
hidden causes of cancer in individual patients that may be missed when analyzing each patient's data in isolation. We believe that this approach can sugge
cancer has not responded to standard therapies. We also believe that the expression data we generate will help other researchers make discoveries that ca

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

With support from

California Initiative to Advance Precision Medicine

St. Baldrick's FOUNDATION

3 Cohorts, 9 Datasets

- Treehouse PED v5 April 2018 (3 datasets)
- Treehouse PED v8 (3 datasets)
- Treehouse public expression dataset (July 2017) (3 datasets)

- Setup by the Treehouse consortium independently
- Fulfill a main goal of the consortium to make the data a public resource
- Manage their own data update and release
- Use the latest Xena Browser without spending engineering resource on it

Treehouse group

Overview of data we visualize

- Functional genomics data
  - "Level 3 data" (in TCGA-speak)

→ No BAMs or FASTQs

- Any clinical, phenotype, or derived data
  - e.g. age, molecular subtype, survival, genomic signature score, computationally derived subgroup, etc

Example analysis: Loading your own data

Steps to viewing your own data:

1. Get or make data
2. Download and install a local Xena Hub
3. Load data into the local Xena Hub
4. Visualize

CGGA (Chinese Glioma Genome Atlas)

Brain tumors datasets over 2,000 samples from Chinese cohorts

https://www.sciencedirect.com/science/article/pii/S1672022921000450

ELSEVIER

Genomics, Proteomics & Bioinformatics

Available online 2 March 2021

In Press, Corrected Proof

Database

Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data from Chinese Glioma Patients

Zheng Zhao [1, #], Ke-Nan Zhang [1, #], Qiangwei Wang [1, 2, #], Guanzhang Li [1], Fan Zeng [1], Ying Zhang [1], Fan Wu [1], Ruichao Chai [1], Zheng Wang [3], Chuanbao Zhang [3], Wei Zhang [3], Zhaoshi Bao [1, 3], Tao Jiang [1, 3, 4, 5]

Step 1: Get data

DataSet ID: mRNAseq_693

Data type: mRNA sequencing

Platform: Illumina HiSeq

Total number of samples: 693

If you use this part of the data (or method included in it), please consider to cite:

1. Wang, Y., Qian, T., You, G., Peng, X., Chen, C., You, Y., Yao, K., Wu, C., Ma, J., Sha, Z., et al. (2015). Localizing seizure-susceptible brain regions associated with low-grade gliomas using voxel-based lesion-symptom mapping. NEURO-ONCOLOGY. 17(2): 282-288.
2. Liu, X., Li, Y., Qian, Z., Sun, Z., Xu, K., Wang, K., Liu, S., Fan, X., Li, S., Zhang, Z., et al. (2018). A radiomic signature as a non-invasive predictor of progression-free survival in patients with lower-grade gliomas. NEUROIMAGE-CLINICAL. 20(1070-1077.

Download

- Clinical Data [Total number of visits: 4705]
- Expression Data from STAR+RSEM [Total number of visits: 4545]
- Raw Fastq Data (BIGD accession number: PRJCA001747)

Step 1: Get data

DataSet ID: mRNAseq_693

Data type: mRNA sequencing

Platform: Illumina HiSeq

Total number of samples: 693

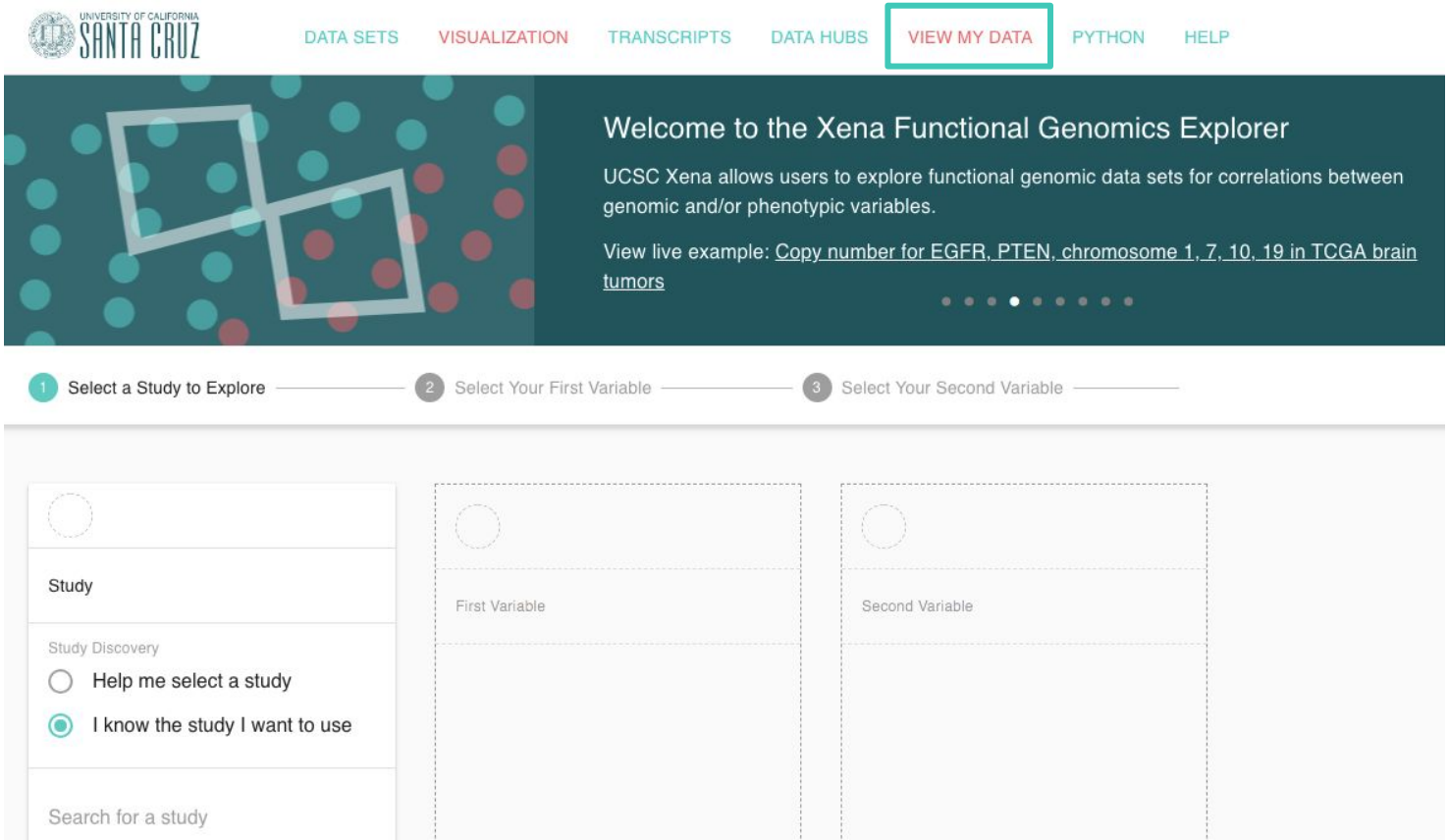
If you use this part of the data (or method included in it), please consider to cite:

1. Wang, Y., Qian, T., You, G., Peng, X., Chen, C., You, Y., Yao, K., Wu, C., Ma, J., Sha, Z., et al. (2015). Localizing seizure-susceptible brain regions associated with low-grade gliomas using voxel-based lesion-symptom mapping. NEURO-ONCOLOGY. 17(2): 282-288.
2. Liu, X., Li, Y., Qian, Z., Sun, Z., Xu, K., Wang, K., Liu, S., Fan, X., Li, S., Zhang, Z., et al. (2018). A radiomic signature as a non-invasive predictor of progression-free survival in patients with lower-grade gliomas. NEUROIMAGE-CLINICAL. 20(1070-1077.

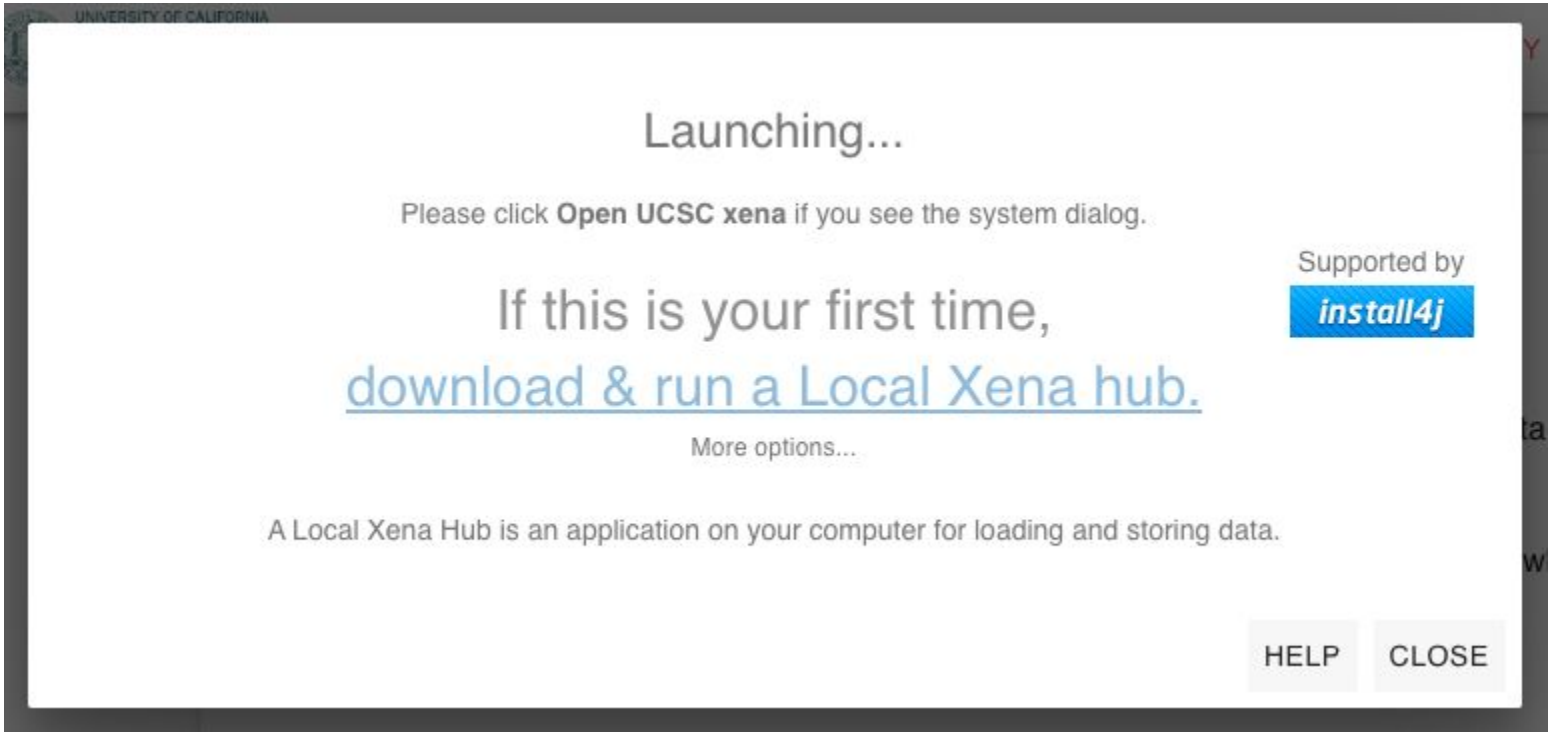
Download

- Clinical Data [Total number of visits: 4705] → **Phenotype data**
- Expression Data from STAR+RSEM [Total number of visits: 4545] → **Expression data**
- Raw Fastq Data (BIGD accession number: PRJCA001747)

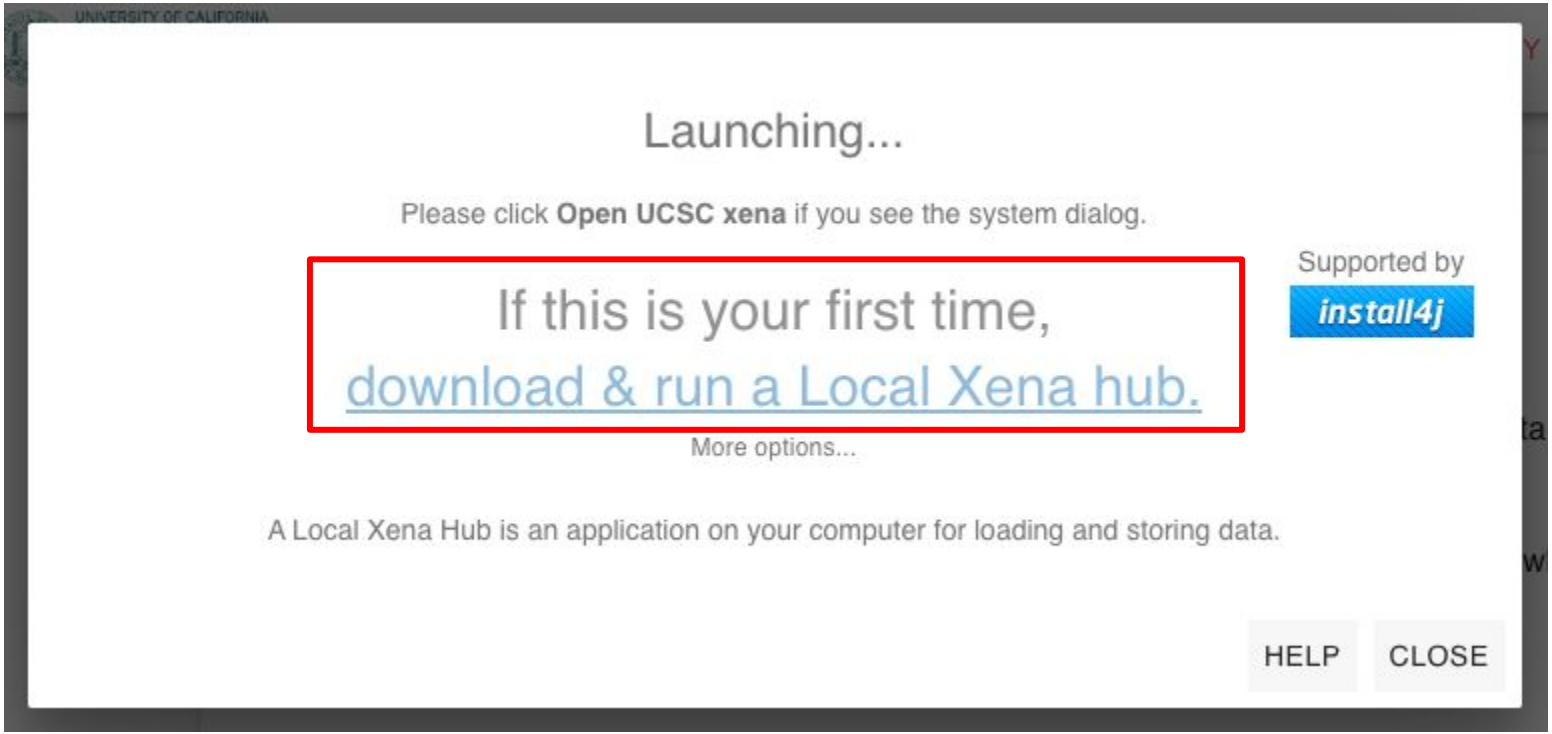
Step 2: Install a Local Hub



Step 2: Install a Local Hub

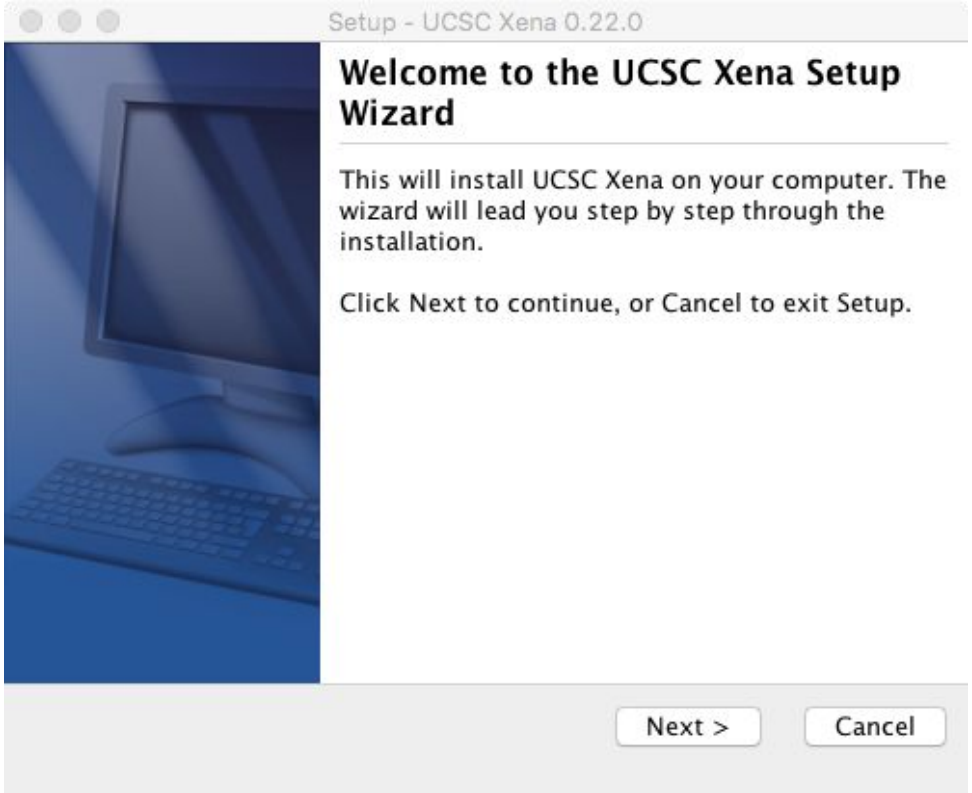


Step 2: Install a Local Hub

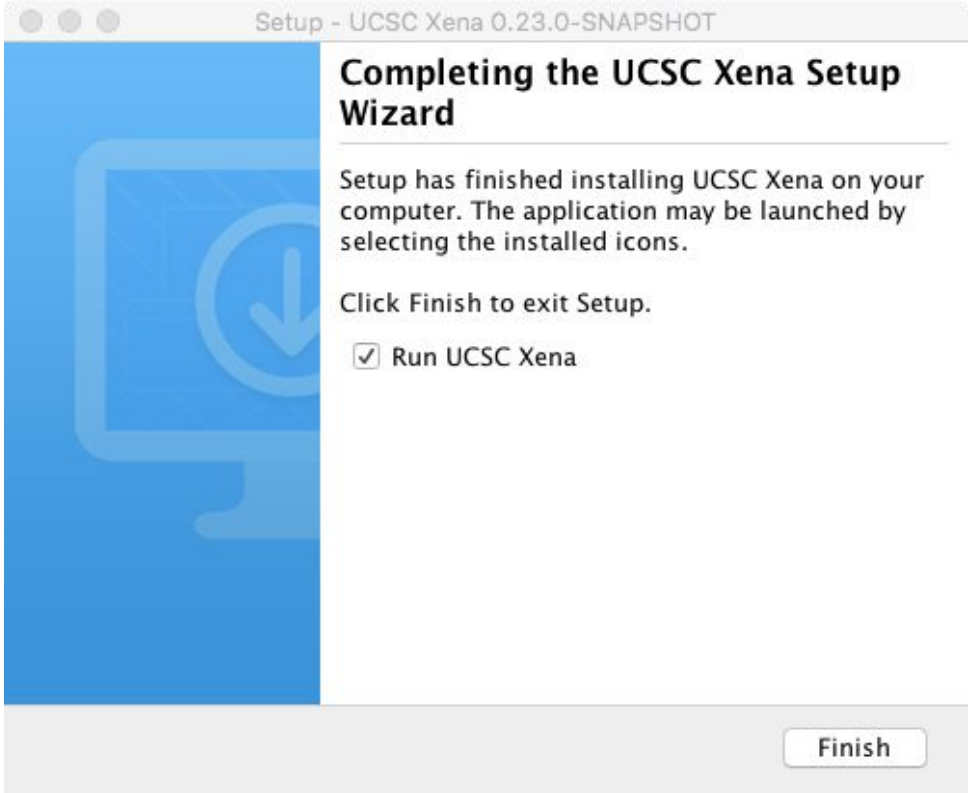


Step 2: Install a Local Hub

Step 2: Install a Local Hub



Step 2: Install a Local Hub



Demo of Step 3 & 4:
Loading data and visualization

Xena's Help Resources

- Tutorials: https://ucsc-xena.gitbook.io/project/tutorials
  - Beginning and Advanced
- Live Examples: https://ucsc-xena.gitbook.io/project/live-examples
- Help: https://ucsc-xena.gitbook.io/project/
- Mailing List: genome-cancer@soe.ucsc.edu
- Public Forum: https://groups.google.com/forum/?fromgroups#!forum/ucsc-cancer-genomics-browser

Logistics for hands on workshop

We will be using breakout rooms so if you will not be attending, please leave right after Q&A.

Thank you!

genome-cancer@soe.ucsc.edu

@UCSCXena

Subscribe to our mailing list:
https://xena.ucsc.edu/#subscribe

We're constantly improving UCSC Xena.
Subscribe to our newsletter and stay up to date.

Your email address | Subscribe

Chan Zuckerberg Initiative

http://xena.ucsc.edu

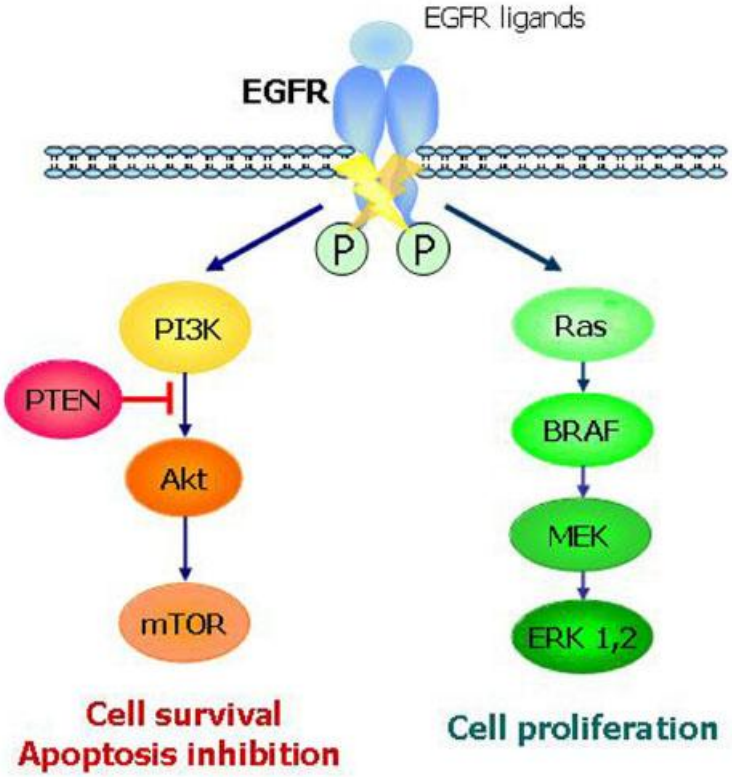
Hands-on Workshop

https://docs.google.com/document/d/1EzSO1HCkQchanLFxcwAf5PmZ3IvHylUXAPGmHR4gpXk

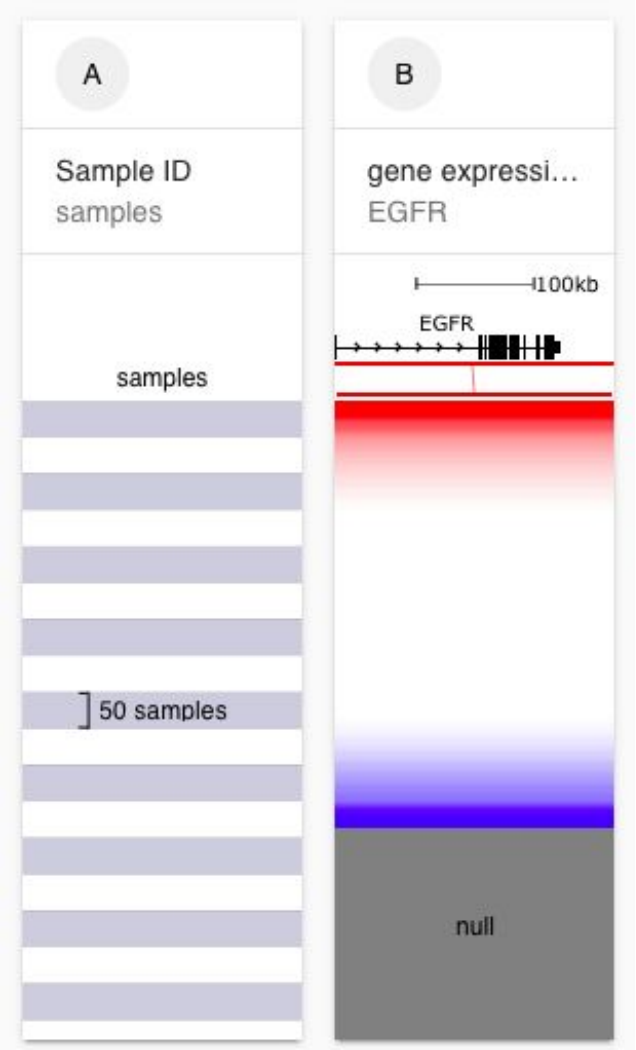
Hands-on Workshop
Tutorial 1: EGFR in Lung Cancer

Overview of *EGFR* in Lung Cancer

- Epidermal Growth Factor Receptor
- *EGFR* aberrations (mutations or amplifications) are present in 10–35% of Lung Adenocarcinoma patients
- *EGFR* aberrations are more common in women

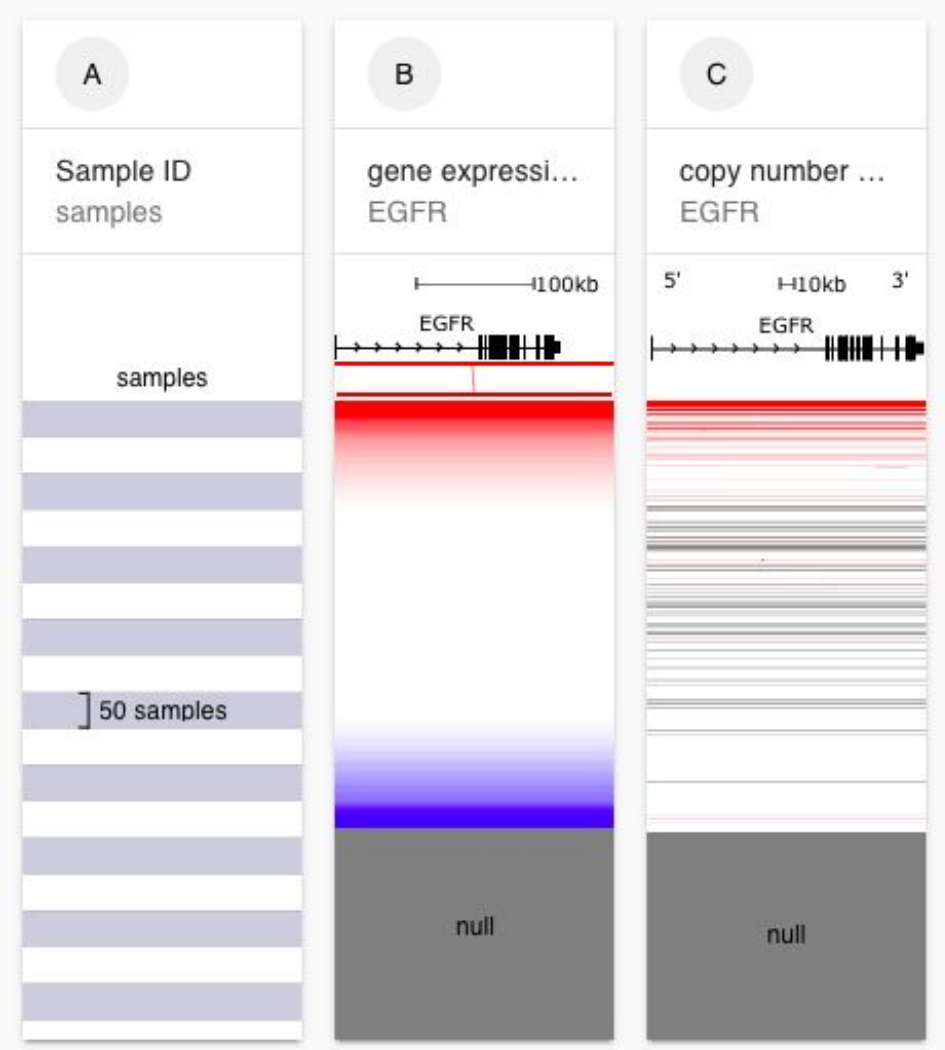


Saletti, et al 2015



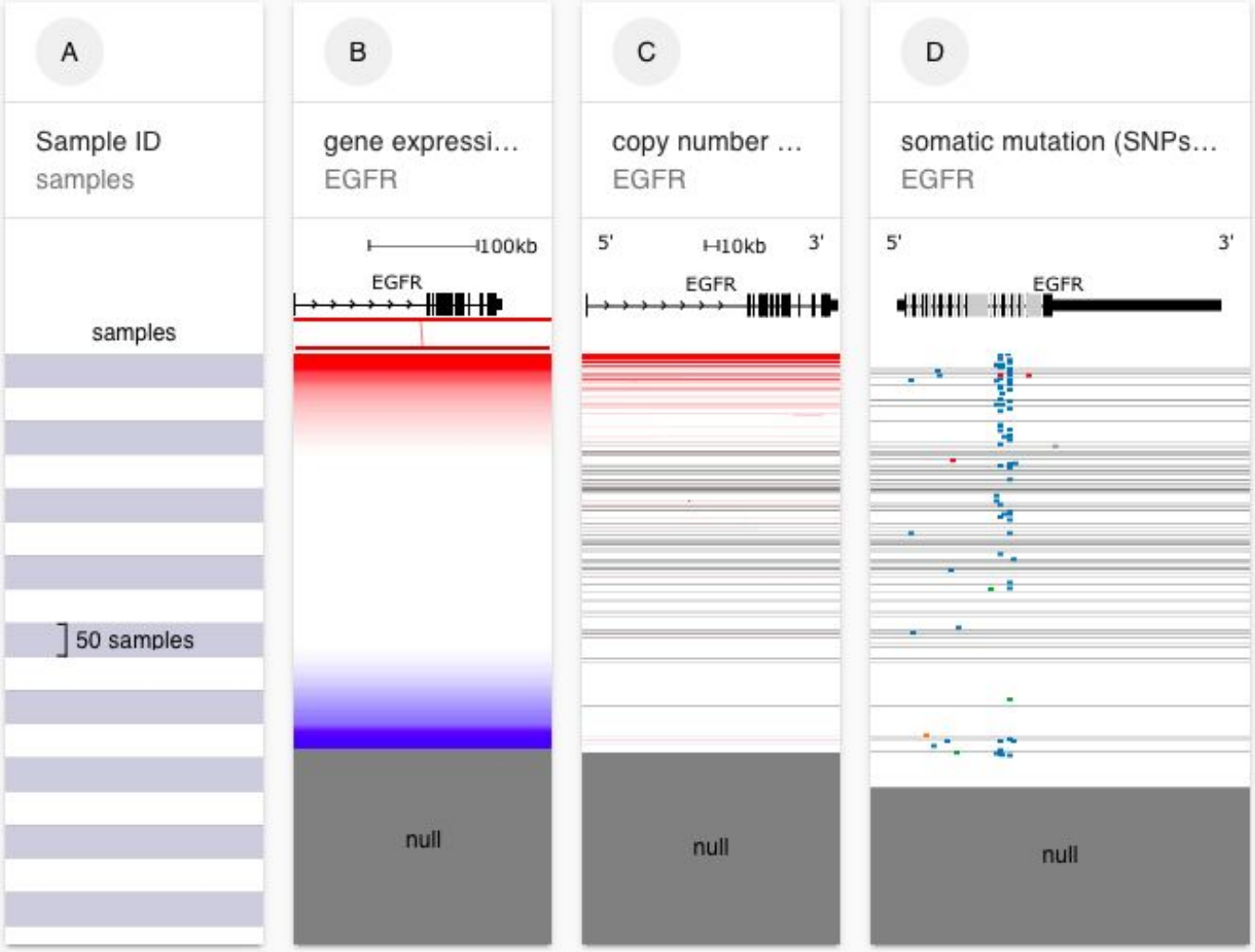
Gene Expression

Some samples have relatively high *EGFR* gene expression



Copy Number Variation

Some samples with higher expression of *EGFR* have an amplification in *EGFR*



Somatic Mutation

Some samples with higher expression of *EGFR* have missense mutations in *EGFR*


A
Sample ID
samples
samples
50 samples
B
gene expressi…
EGFR
100kb
EGFR
null
C
copy number …
EGFR
5'
10kb
3'
EGFR
null
D
somatic mutation (SNPs…
EGFR
5'
3'
EGFR
null

Advanced Tutorial 2: PAM50 breast cancer subtypes

PAM50 Breast Cancer Subtypes

PAM50 is a list of 50 genes that classifies breast cancer into five molecular intrinsic subtypes: Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like.

- Each molecular subtype has a different prognosis

→ We are going to look at these 50 genes in relationship to the subtype calls

TFAC30 gene expression signature

Gene expression signatures are a mathematical formula performed over a group of genes

- Gives a single number that summarizes gene expression of many genes
- Can be used to predict response to therapy (e.g. a high value means you are likely to respond)

→ We are going to look at the TFAC30 gene expression signature

- Signature over 30 genes predicts pCR to (T/FAC) chemo

Thank you!

genome-cancer@soe.ucsc.edu

@UCSCXena

Subscribe to our mailing list:
https://xena.ucsc.edu/#subscribe

We're constantly improving UCSC Xena.
Subscribe to our newsletter and stay up to date.

Your email address | Subscribe

Chan Zuckerberg Initiative

http://xena.ucsc.edu