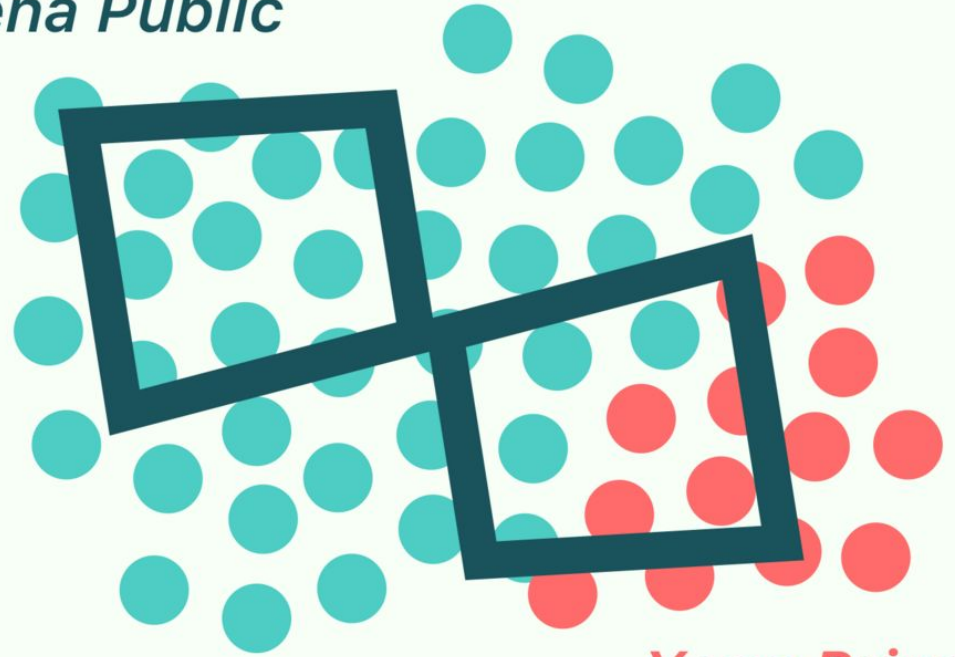


UCSC Xena
**See the
bigger
picture**

Xena Public



Xena Private

Jingchun Zhu, David Haussler

University of California Santa Cruz Genomics Institute
ITCR Annual Meeting, June 1, 2017

Outline

Xena overview

New visualizations

New browser features

Integration with other tools

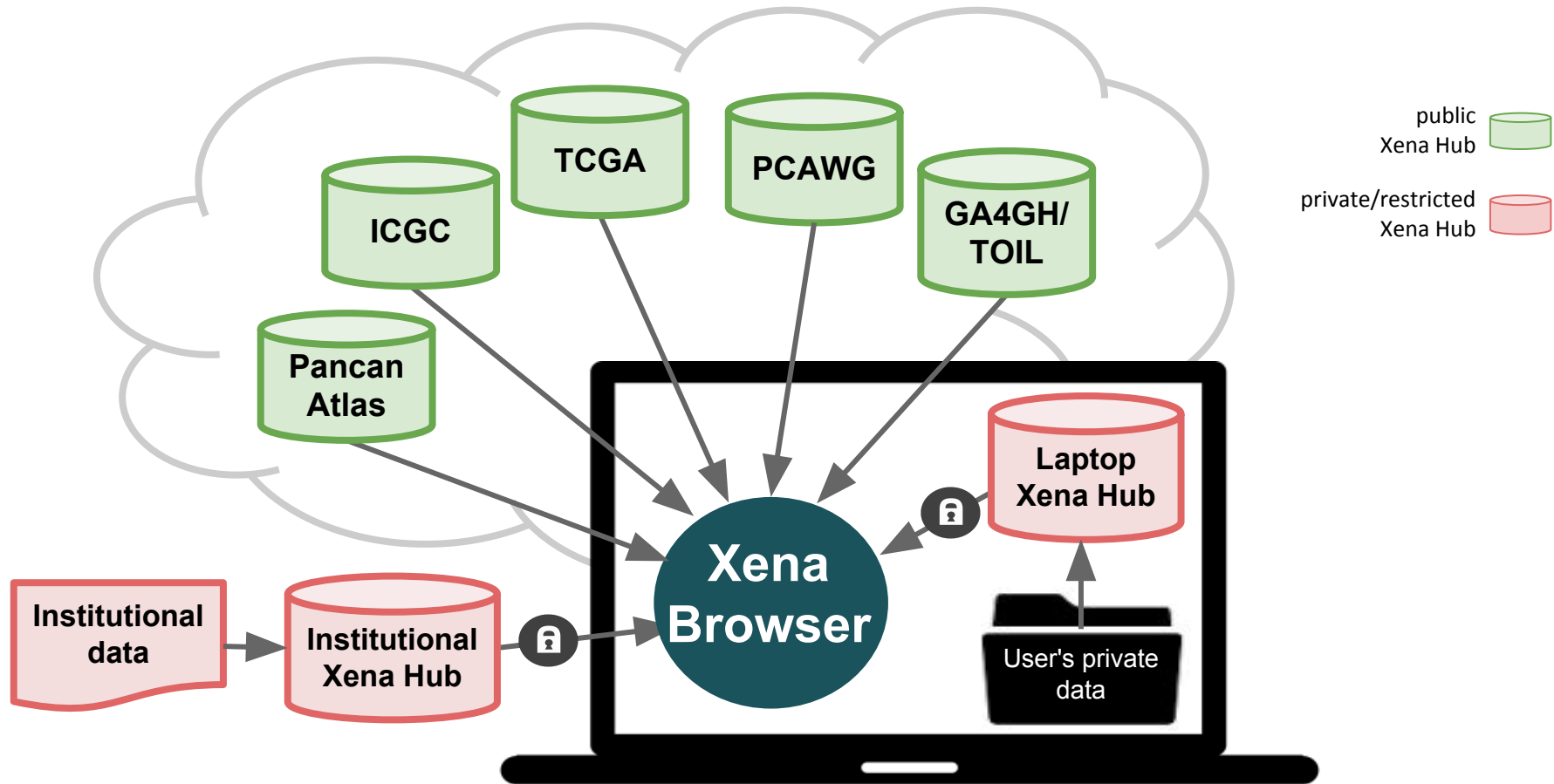
Future work

Xena overview

**We need a shared and standardized
global network of genomic data**



Both of large repositories and small studies



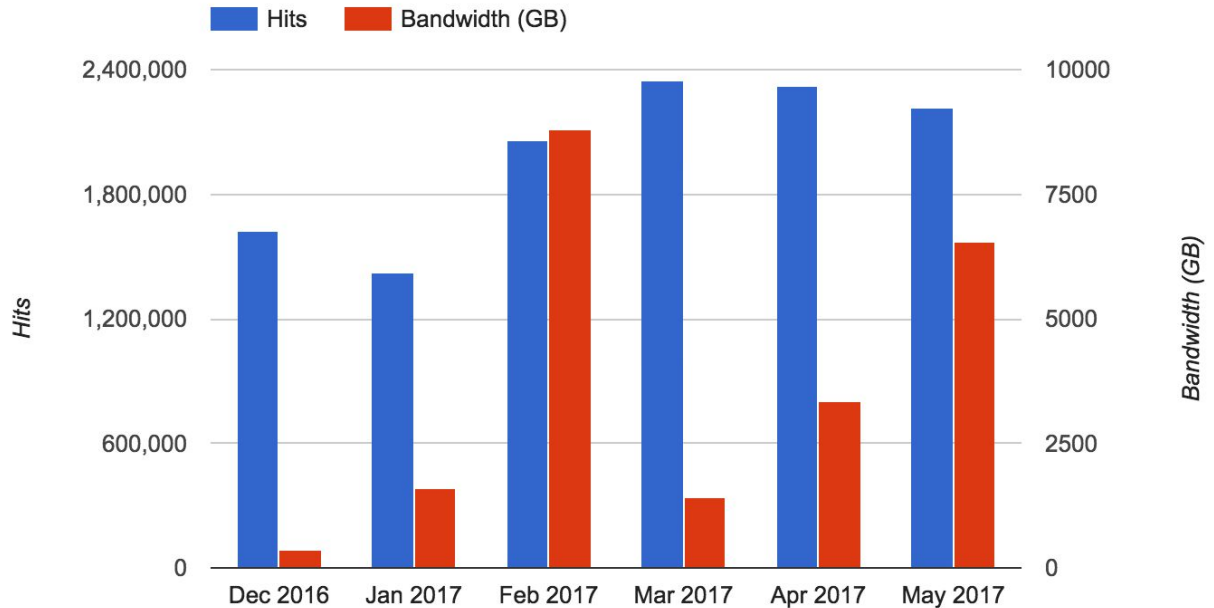
- Federated data hubs
- Data combined in the browser, ensuring data security

Data types Xena visualizes

- SNPs and small INDELS
- Structural variants
- Segmented copy number, gene-level copy number
- Gene-, Transcript-, Exon-, Protein-, and miRNA-expression
- DNA methylation (genes and probes)
- Phenotype, clinical data
- Signature scores, classifications, derived parameters

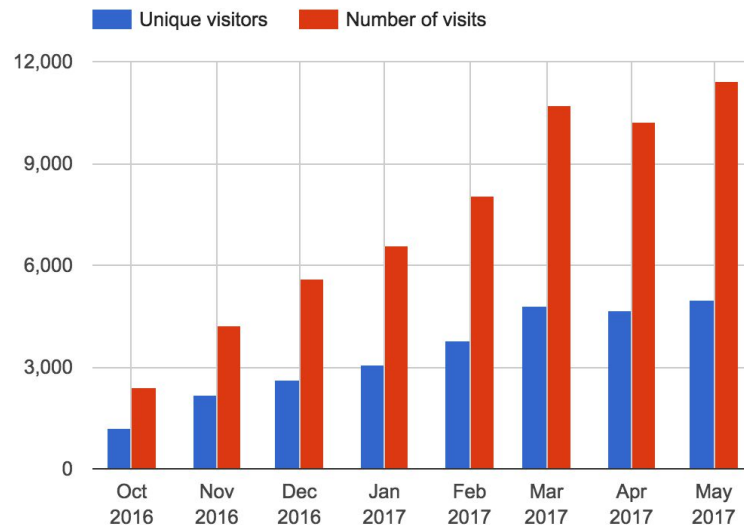
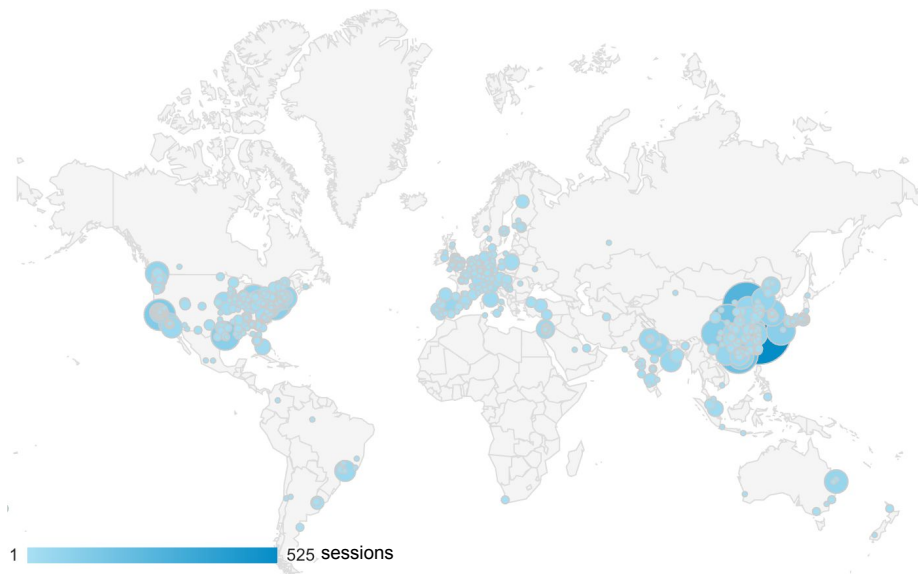
Xena Hub Usage

- Millions of hits per month
- TB data visualized/downloaded
- 430 laptop hubs 'phoned home' in April 2017



Xena Browser usage

- 6,915 sessions in the past month
- Average 06:56 min per session



New Whole Genome Data Visualizations

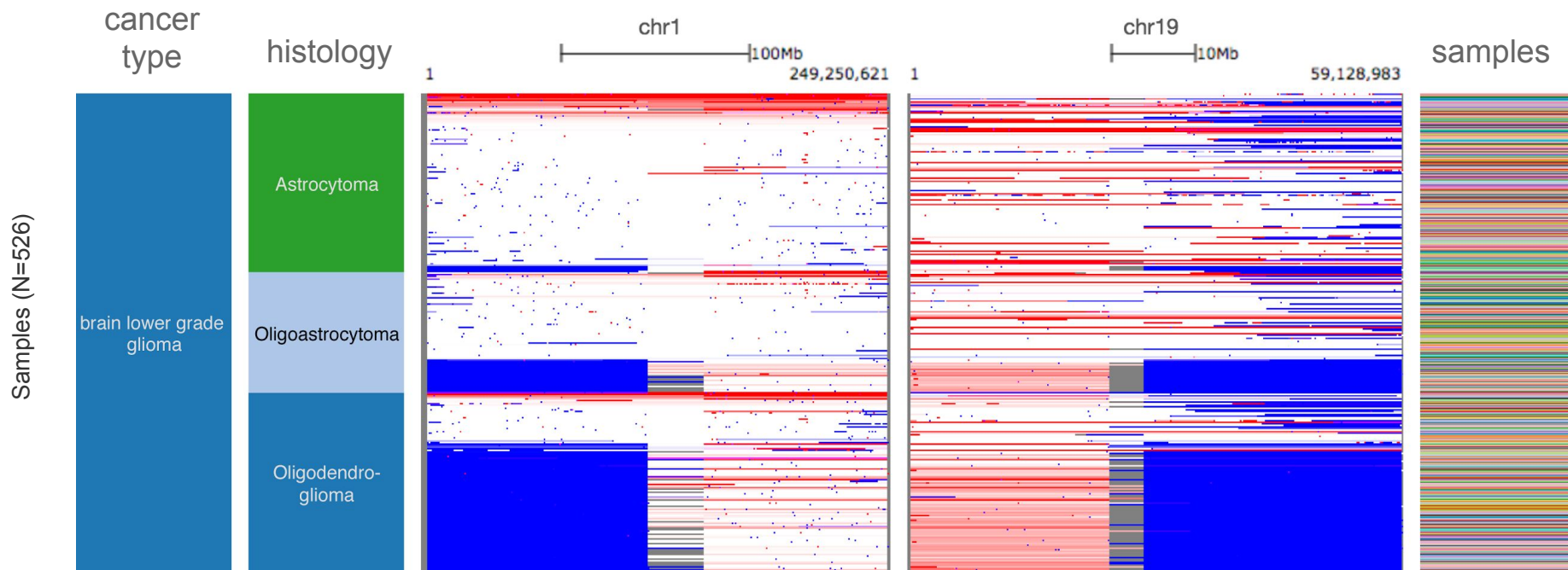
Driven by the **PCAWG Project**

- Coding & non-coding regions
- Gene- & coordinate-centric views
- Copy number variations
- Simple mutations
- Structural variants

Query examples:

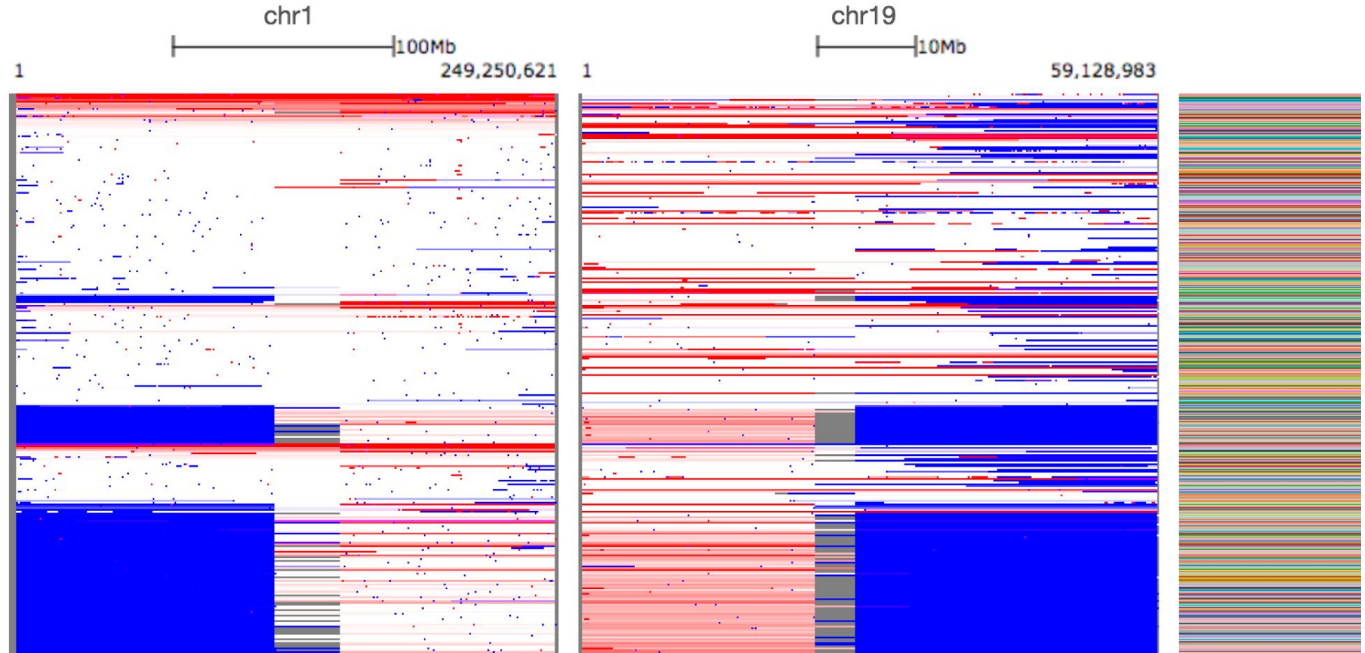
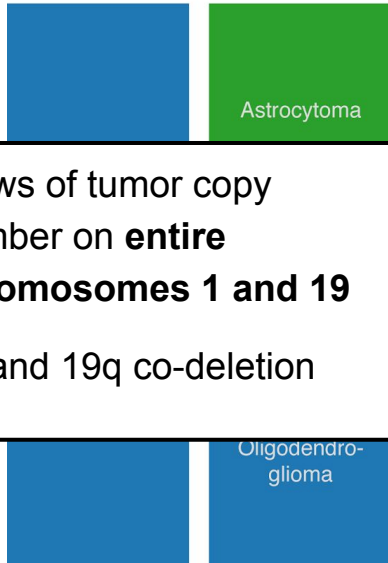
'chr1' 'chr19q' 'chr21:42870119-42870526' 'TP53'

1p/19q Co-deletion in lower grade glioma



1p/19q Co-deletion in lower grade glioma

Views of tumor copy number on **entire chromosomes 1 and 19**
1p and 19q co-deletion

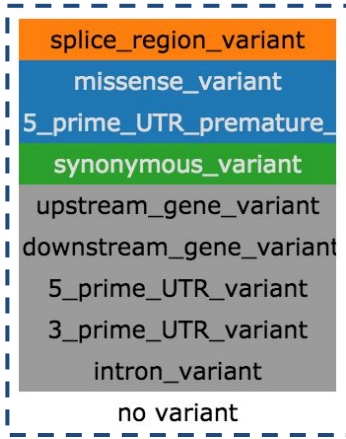
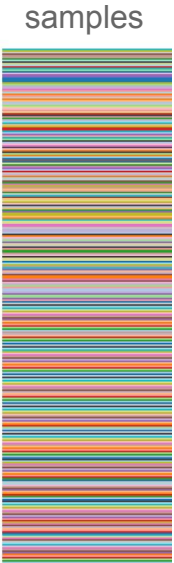
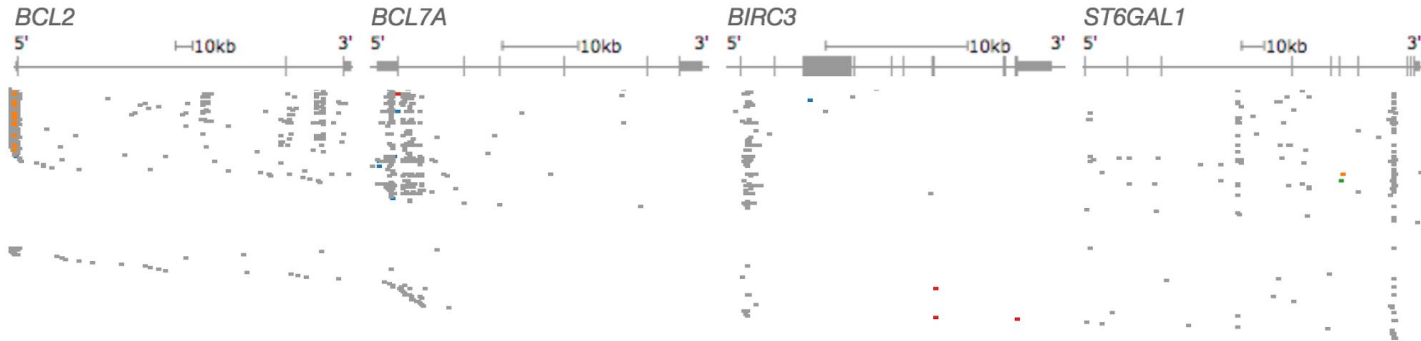


Intron Mutations in Lymphoma genes

Samples (N=321)

Malignant Lymphoma : Germinal center B-cell derived lymphomas

Chronic Lymphocytic Leukemia : CLL with mutated and unmutated IgVH



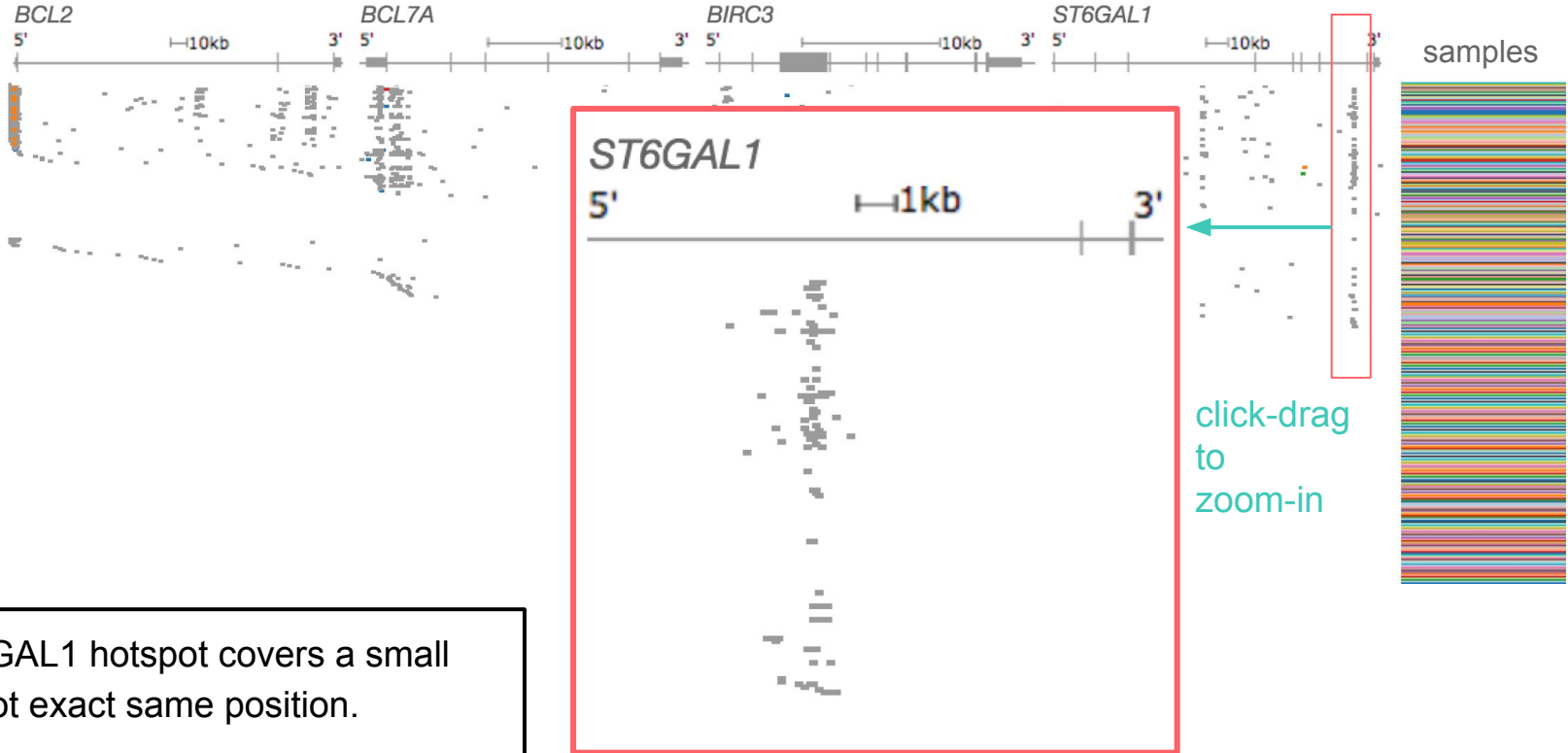
Intron mutation hotspots in lymphoma
Hotspots overlapping enhancer regions

Intron Mutations in Lymphoma genes

Samples (N=321)

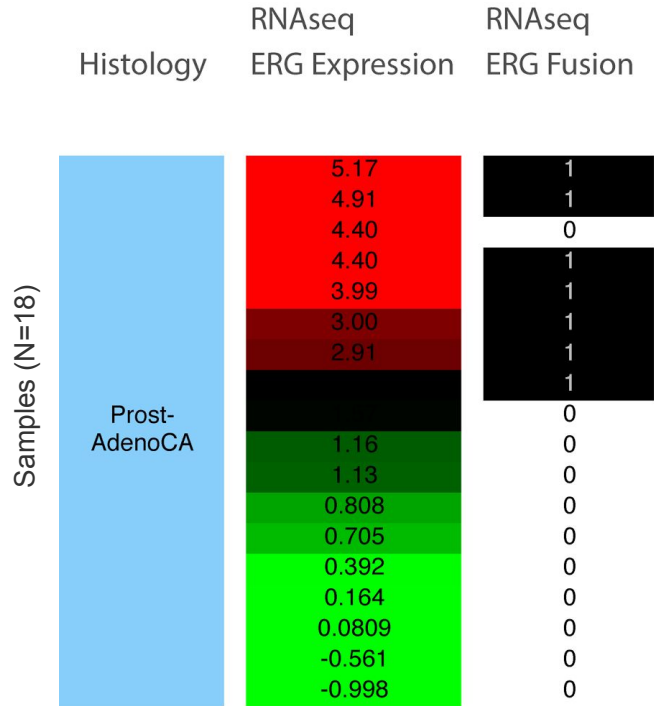
Malignant Lymphoma : Germinal center B-cell derived lymphomas

Chronic Lymphocytic Leukemia : CLL with mutated and unmutated IgVH



The ST6GAL1 hotspot covers a small region, not exact same position.

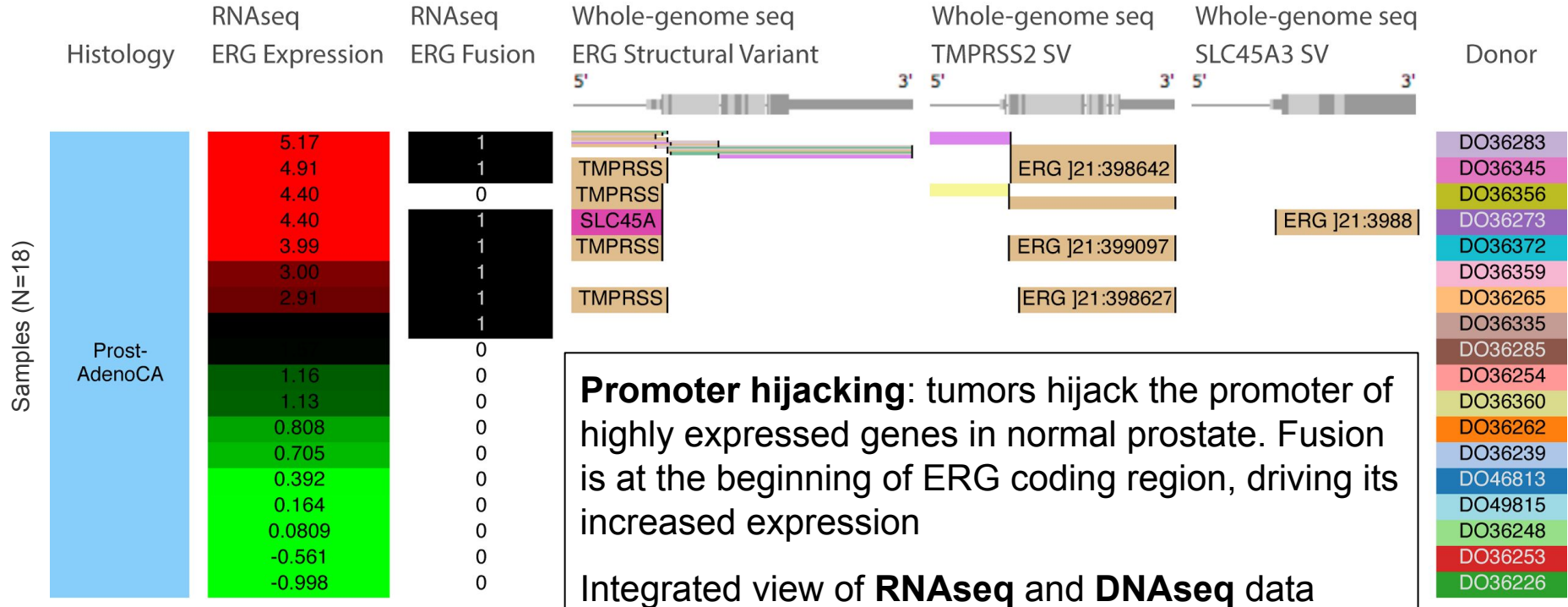
Structural Variant: ERG Fusion in Prostate Cancer



Recurrent fusion-driven **ERG** overexpression drives a subset of prostate cancer

Displaying **RNAseq** analysis results

Structural Variant: ERG Fusion in Prostate Cancer



New Browser Features

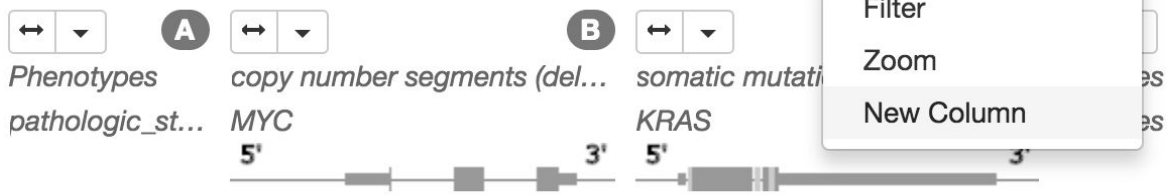
Search, Filter and Group samples

Powerful, text-based search interface

- Simple search:
 - 'Make two groups of samples: those that have a ATRX missense mutation and those that do not' => 'missense'
- Complicated search:
 - 'Show me all samples that are stage III or IV *with* MYC amplification or KRAS mutation' => '*(stage III OR stage IV) AND (B:>0.5 OR KRAS)*'

(stage III OR stage IV) AND (B:>0.5 OR KRAS)

Matching samples: 51 Help with search



Samples (N=480)



Other New Browser Features

- Adjustable KM plot time axis: e.g. 1, 3, 5 years
- Coloring in linear or log scale
- Statistics for box plot, scatter plot and KM plot
- Enhanced PDF
- Download the entire spreadsheet at once
- Bookmarks

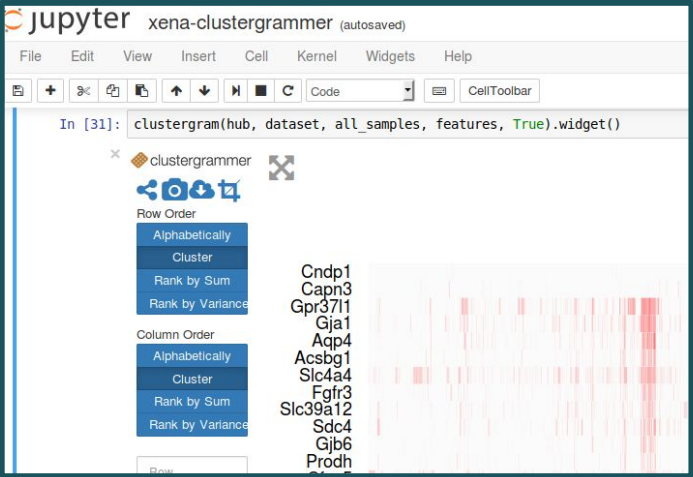
Integration with other tools

Jupyter Notebook/Python API

The screenshot shows a Jupyter Notebook interface with the following elements:

- Header:** Jupyter logo and text "xena-clustergrammer (autosaved)".
- Menu Bar:** File, Edit, View, Insert, Cell, Kernel, Widgets, Help.
- Toolbar:** Includes icons for file operations, a dropdown menu set to "Code", and a "CellToolbar" button.
- Code Cell:** Contains the Python code: `clustergram(hub, dataset, all_samples, features, True).widget()`.
- Widget Output:** A clustergram visualization titled "clustergrammer". It features:
 - Row Order Controls:** A vertical list of buttons: "Alphabetically", "Cluster", "Rank by Sum", and "Rank by Variance".
 - Column Order Controls:** A vertical list of buttons: "Alphabetically", "Cluster", "Rank by Sum", and "Rank by Variance".
 - Gene Labels:** A list of gene names: Cndp1, Capn3, Gpr3711, Gja1, Aqp4, Acsbg1, Slc4a4, Fgfr3, Slc39a12, Sdc4, Gjb6, and Prodh.
 - Heatmap:** A grid of red and white cells representing data values for each gene across multiple samples.

Jupyter Notebook/Python API



The screenshot shows a Jupyter Notebook window titled "xena-clustergrammer (autosaved)". The code cell contains the following Python code:

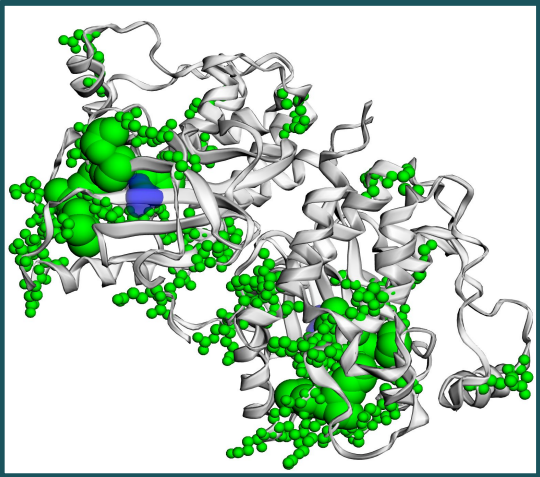
```
In [31]: clustergram(hub, dataset, all_samples, features, True).widget()
```

The output is a "clustergrammer" widget. It includes a "Row Order" section with buttons for "Alphabetically", "Cluster", "Rank by Sum", and "Rank by Variance". The "Column Order" section has buttons for "Alphabetically", "Cluster", "Rank by Sum", and "Rank by Variance". A list of gene names is displayed on the right side of the widget:

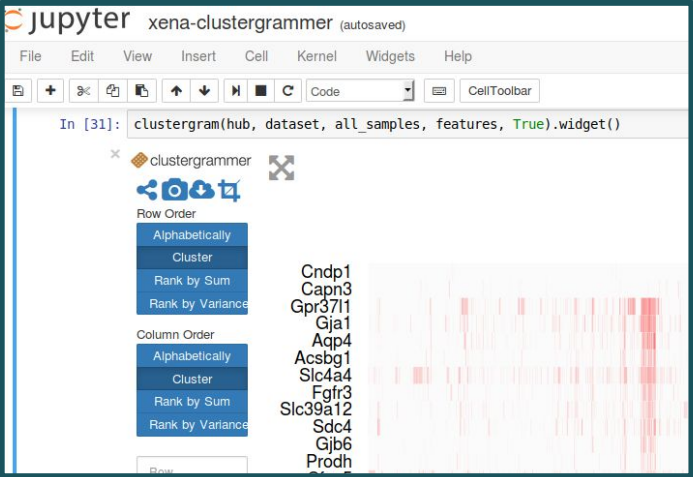
- Cndp1
- Capn3
- Gpr37l1
- Gja1
- Aqp4
- Acsbg1
- Slc4a4
- Fgfr3
- Slc39a12
- Sdc4
- Gjb6
- Prodh

Below the gene names is a heatmap visualization with red and white cells.

MuPIT



Jupyter Notebook/Python API

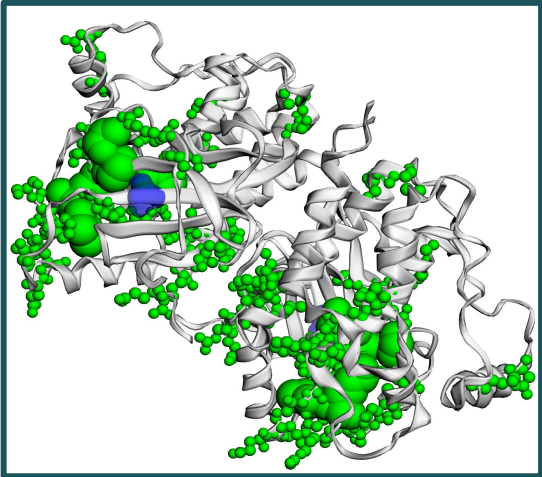


The screenshot shows a Jupyter Notebook window titled "xena-clustergrammer (autosaved)". The code cell contains the following Python code:

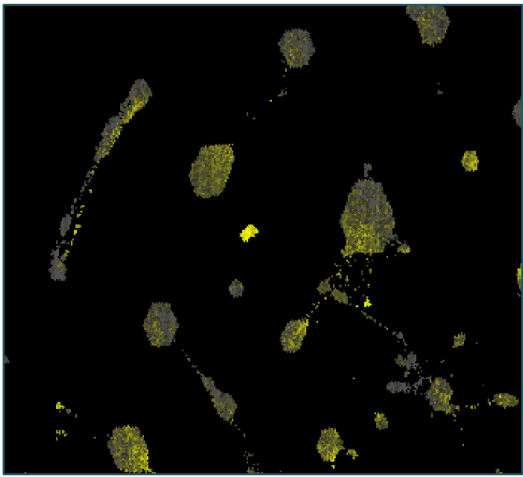
```
In [31]: clustergram(hub, dataset, all_samples, features, True).widget()
```

Below the code, a widget titled "clustergrammer" is displayed. It includes a "Row Order" section with buttons for "Alphabetically", "Cluster", "Rank by Sum", and "Rank by Variance". The "Column Order" section has buttons for "Alphabetically", "Cluster", "Rank by Sum", and "Rank by Variance". To the right of these buttons is a list of gene names: Cndp1, Capn3, Gpr3711, Gja1, Aqp4, Acsgb1, Slc4a4, Fgfr3, Slc39a12, Sdc4, Gjb6, and Prodh. A heatmap visualization is partially visible on the right side of the widget.

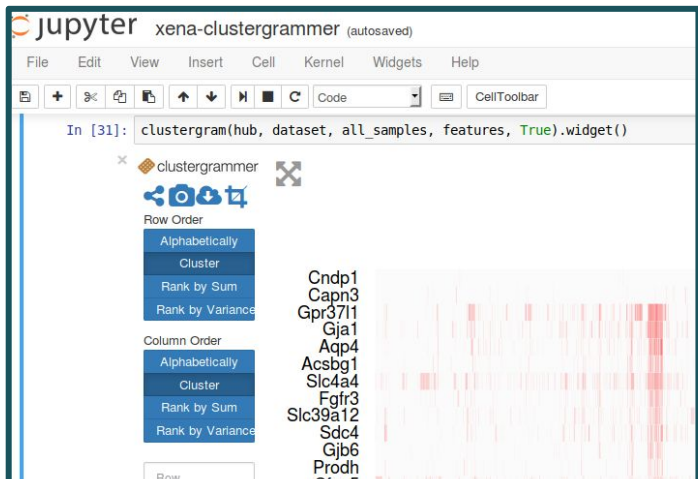
MuPIT



TumorMap



Jupyter Notebook/Python API

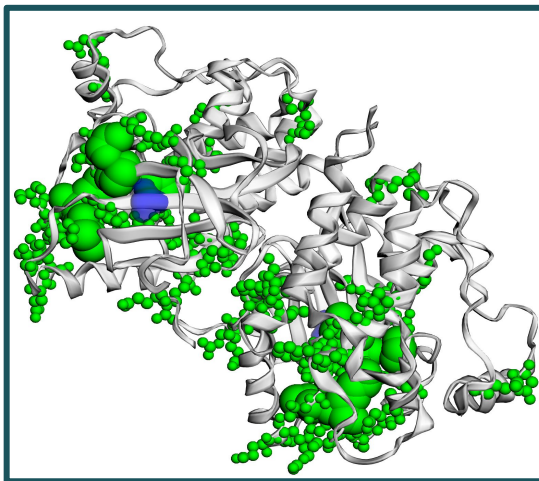


The screenshot shows a Jupyter Notebook window titled "xena-clustergrammer (autosaved)". The code cell contains the following Python code:

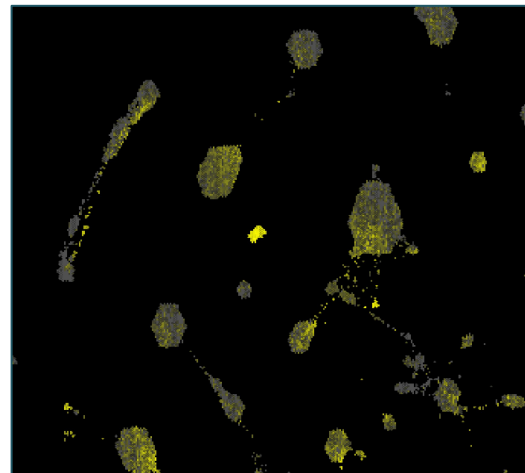
```
In [31]: clustergram(hub, dataset, all_samples, features, True).widget()
```

Below the code cell, a widget for "clustergram" is displayed. It includes a "Row Order" section with buttons for "Alphabetically", "Cluster", "Rank by Sum", and "Rank by Variance". The "Column Order" section has similar buttons. A list of gene symbols is shown, including Cndp1, Capn3, Gpr37l1, Gja1, Aqp4, Acsbg1, Slc4a4, Fgfr3, Slc39a12, Sdc4, Gjb6, and Prodh. A heatmap visualization is partially visible on the right side of the widget.

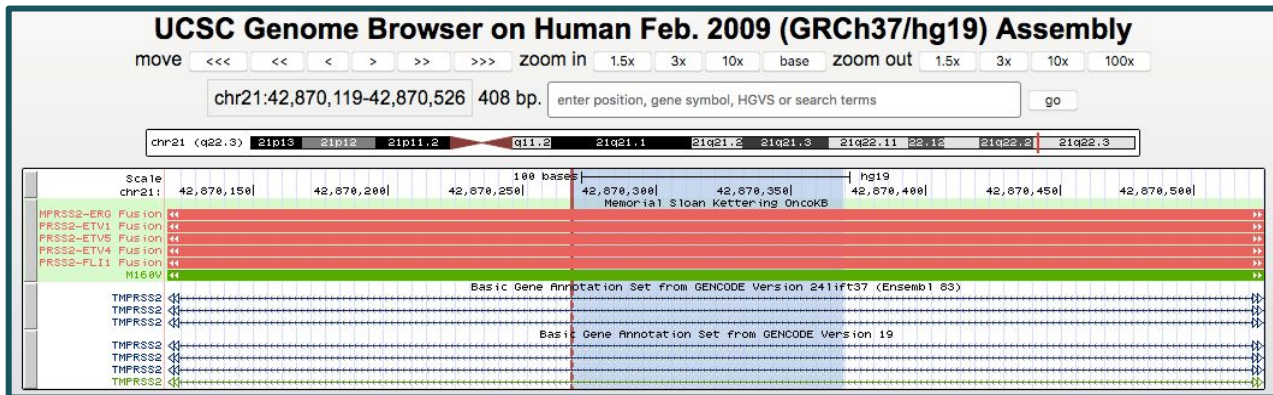
MuPIT



TumorMap



UCSC Genome Browser

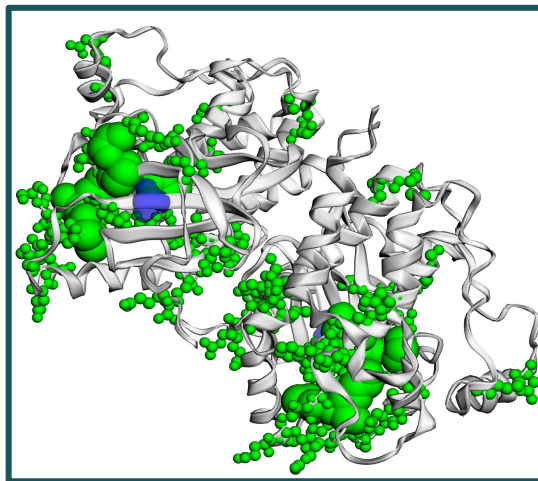


The screenshot shows the UCSC Genome Browser interface for human chromosome 21. The title is "UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly". The search bar shows "chr21:42,870,119-42,870,526 408 bp." and "enter position, gene symbol, HGVS or search terms". The browser displays a genomic track for chromosome 21 (q22.3) with various annotations. The track includes a scale bar, a gene annotation set from GENCODE Version 2411ft37 (Ensembl 63), and a gene annotation set from GENCODE Version 19. The track shows several genes, including MPRSS2-ERG, PRSS2-ETV1, PRSS2-ETV5, PRSS2-ETV4, PRSS2-FL11, and M160V. The track also shows the location of the Memorial Sloan Kettering OncokB. The track is zoomed in to a 100x scale, showing a 100 base region from 42,870,150 to 42,870,500.

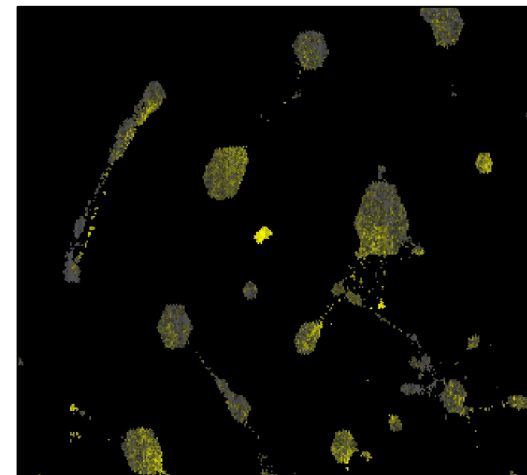
Jupyter Notebook/Python API

The screenshot shows a Jupyter Notebook window titled "xena-clustergrammer (autosaved)". The code cell contains the following Python code: `clustergram(hub, dataset, all_samples, features, True).widget()`. Below the code, a widget titled "clustergram" is displayed. It includes a "Row Order" section with buttons for "Alphabetically", "Cluster", "Rank by Sum", and "Rank by Variance". A "Column Order" section has similar buttons. To the right, a list of gene symbols is shown: Cndp1, Capn3, Gpr37l1, Gja1, Aqp4, Acsbg1, Slc4a4, Fgfr3, Slc39a12, Sdc4, Gjb6, and Prodh. A heatmap visualization is partially visible on the right side of the widget.

MuPIT



TumorMap



UCSC Genome Browser

The screenshot shows the UCSC Genome Browser interface for human chromosome 21. The title is "UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly". The current view is centered on the region chr21:42,870,119-42,870,526 (408 bp). The interface includes navigation buttons (move, zoom in, zoom out) and a search bar. Below the search bar, a track shows the chromosome structure with bands for 21q13, 21q12, 21q11.2, q11.2, 21q21.1, 21q21.2, 21q21.3, 21q22.11, 22.12, 21q22.2, and 21q22.3. A scale bar indicates 100 bases. The main track displays various genomic features, including PRSS2-ERG, PRSS2-ETV1, PRSS2-ETV5, PRSS2-ETV4, PRSS2-FL11, M169V, and TMFRSS2. The TMFRSS2 track shows multiple gene models with exons and introns.

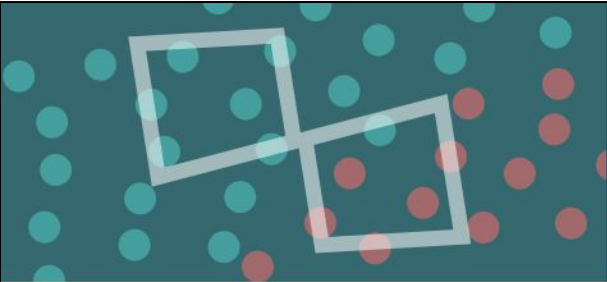
Genomic Data Commons

data coming soon!

In-progress ...

“Building a Spreadsheet” Wizard

Welcome screen / select study



Welcome to the Xena Functional Genomics Explorer ✕

UCSC Xena allows users to explore functional genomic data sets for correlations between genomic and/or phenotypic variables.

View live example: [TP53 Expression vs. Mutation in TCGA Pan-Cancer](#)

1 Select a Study to Explore

2 Select Your First Variable

3 Select Your Second Variable

Study

If you would like help determining the data set to use, Xena can suggest data sets if you provide a primary disease or tissue of origin.

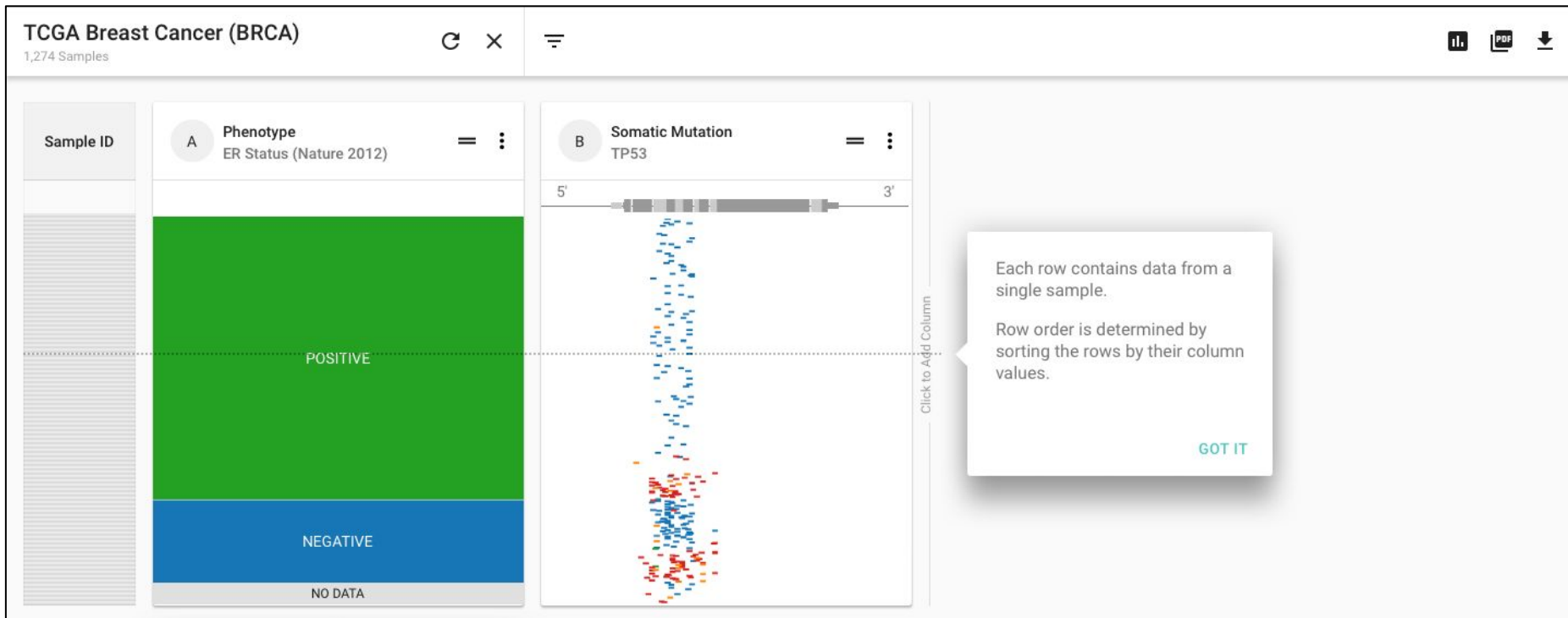
Study

- Help me select a study
- I know the study I want to use

First Variable

Second Variable

Displaying two variables / Done



New adventure ...

Single Cell RNAseq Visualization

1.3 million mouse brain cells

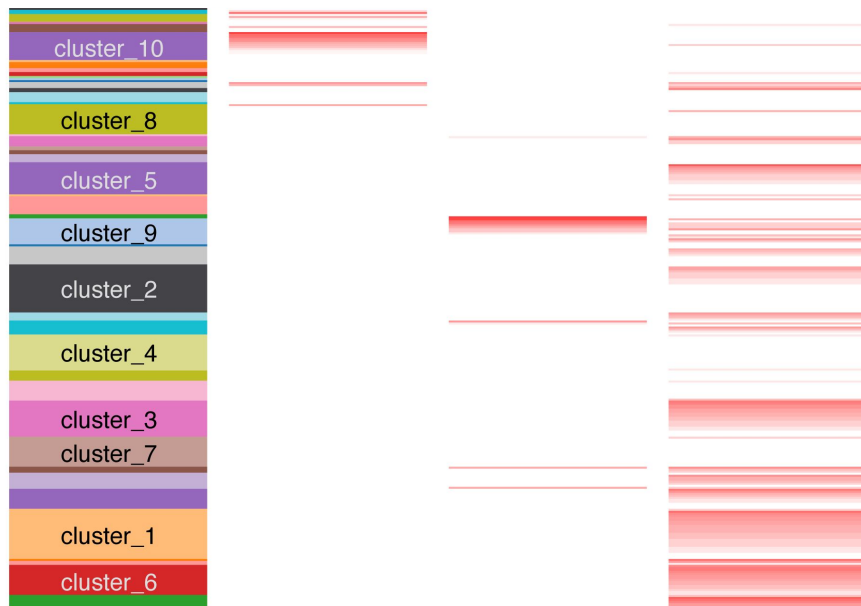
A 1M neurons s... *ALDOC*
Graph_cluste...

B 1M neurons s... *Crym*

C 1M neurons s... *Gria2*

D

Samples (N=1,306,127)

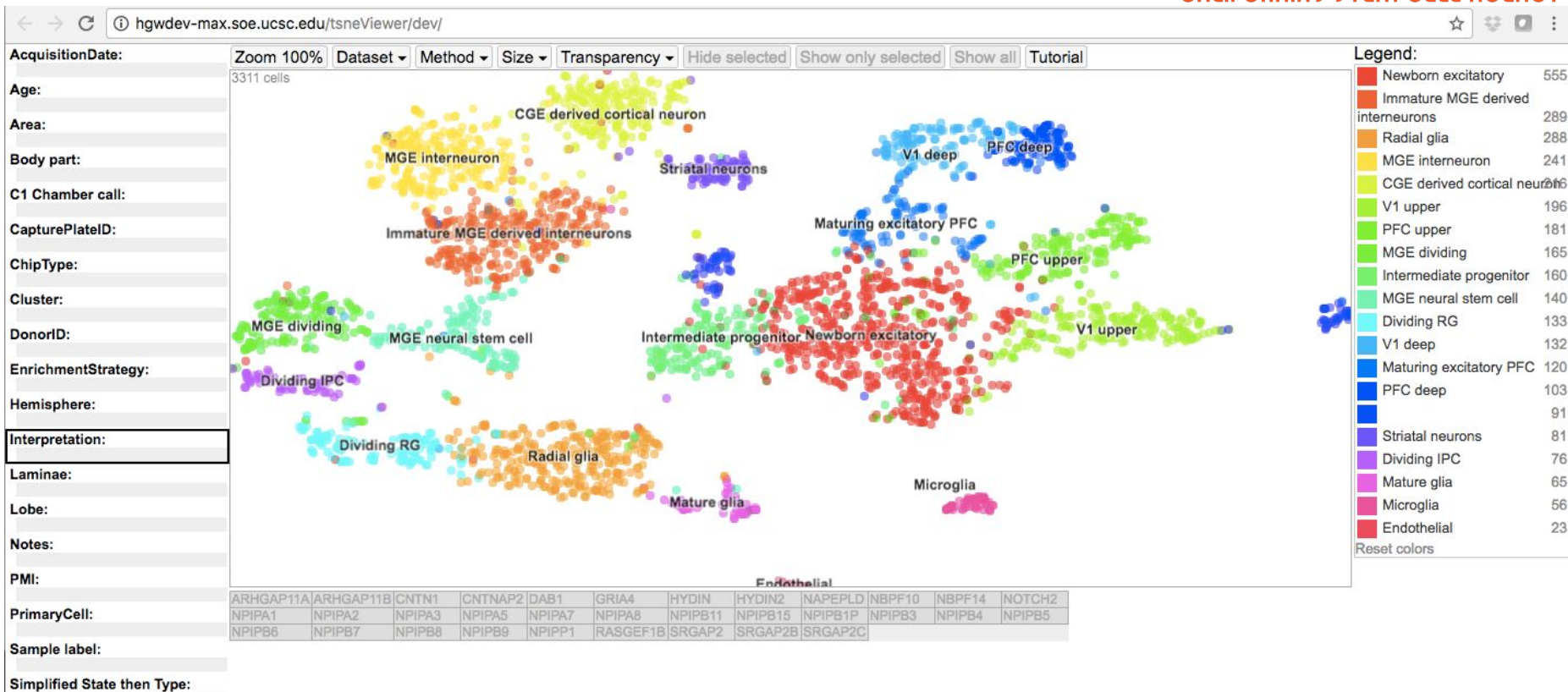


1.3 million embryonic mouse brain cells
single-cell RNAseq

10x genomics
demonstration dataset

Stem Cell Genomics

Kent, Haussler, Salama, Kriegstein, Pollen, Nowakowski



Funding

National Cancer Institute
Amazon Web Services



UNIVERSITY OF CALIFORNIA
SANTA CRUZ

Genomics
Institute

Collaborators

TOIL Team
UCSC Genomics Core
UCSC Genome Browser
PCAWG consortium
ITOMIC trial
Treehouse
Angela Brooks's Lab
Eric Collison
Tumor Map - Stuart Lab
MuPIT/CRAVAT
BioJS



PCAWG
PanCancer Analysis
OF WHOLE GENOMES



Big Data to
Knowledge(BD2K)



Global Alliance
for Genomics & Health



BIOJS



NATIONAL CANCER INSTITUTE
Informatics Technology for
Cancer Research



<http://xena.ucsc.edu>

<https://github.com/ucscXena>

Jing Zhu jzhu@soe.ucsc.edu

Mary Goldman mary@soe.ucsc.edu

Brian Craft craft@soe.ucsc.edu