The UCSC Xena platform for public and private cancer genomics data visualization and interpretation

Mary Goldman[1+], Brian Craft[1+], Mim Hastie[2], Kristupas Repečka[3], Akhil Kamath[4], Fran McDade[2], Dave Rogers[2], Angela N. Brooks[5], Jingchun Zhu[1*], and David Haussler[1]

[1] UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA

[2] Clever Canary, New York, NY, USA

[3] Vilnius University, Vilnius, Lithuania

[4] Birla Institute of Technology and Science, Goa, India

[5] Department of Biomolecular Engineering, University of California, Santa Cruz, CA, USA

+ These authors contributed equally to this work

* Corresponding author

Abstract

UCSC Xena is a visual exploration resource for both public and private omics data, supported through the web-based Xena Browser and multiple turn-key Xena Hubs. This unique archecture allows researchers to view their own data securely, using private Xena Hubs, simultaneously visualizing large public cancer genomics datasets, including TCGA and the GDC. Data integration occurs only within the Xena Browser, keeping private data private. Xena supports virtually any functional genomics data, including SNVs, INDELs, large structural variants, CNV, expression, DNA methylation, ATAC-seq signals, and phenotypic annotations. Browser features include the Visual Spreadsheet, survival analyses, powerful filtering and subgrouping, statistical analyses, genomic signatures, and bookmarks. Xena differentiates itself from other genomics tools, including its predecessor, the UCSC Cancer Genomics Browser, by its ability to easily and securely view public and private data, its high performance, its broad data type support, and many unique features.

Introduction

There is a great need for easy-to-use cancer genomics visualization tools for both large public data resources such as TCGA (The Cancer Genome Atlas) (Chin 2011, Chin 2011) and the GDC (Genomic Data Commons) (Grossman 2016) as well as smaller-scale datasets generated by individual labs. It is important that visualization and exploration tools put these large, complex datasets in the hands of researchers who may not have computational expertise. Commonly-used interactive visualization tools are either web-based portals or desktop applications. Data portals have a dedicated backend and are a powerful means of viewing centrally-hosted resource datasets. Examples of these data portals include the UCSC Cancer Genomics Browser (Zhu 2010), cBioPortal (Gao 2013), ICGC Data Portal (https://dcc.icgc.org), GDC Data Portal (Grossman 2016), and FireBrowse (http://firebrowse.org). However, using these data portals to explore an investigator's own private data involves either a redeployment of the entire platform, a difficult task even for bioinformaticians, or uploading private data to a server outside the user's control, a non-starter for protected patient data such as germline variants (e.g. MAGI (Leiserson 2015.), WebMeV (Wang 2017), Ordino (Streit 2018)). Desktop tools can view a user's own data securely (e.g. IGV (Robinson 2011), Gitools (Perez-Llamas 2011)), but lacks well-maintained, prebuilt files for the ever-evolving and expanding public data resources. This dichotomy between data portals and desktop tools highlights the challenge of using a single platform for both large public data as well as smaller-scale datasets generated by individual labs.

Complicating this dichotomy is the increase in cancer genomics data, both public resources and individual datasets, which has been largely due to recent technological advances, including lower-cost high-throughput sequencing (Mardis 2008) as well as single-cell based technologies (Wills 2015). In addition to larger numbers of samples and cells being sequenced, cancer genomics data is

also becoming increasingly complex. For example, in addition to the relatively common gene expression and somatic mutation data, cancer genomics datasets are beginning to be generated using whole-genome sequencing (PCAWG citation), DNA methylation whole genome bisulfite sequencing (Zhou 2018), and nucleosome occupancy assayed by ATAC-seq (Corces 2018). Visualizing and exploring these diverse data modalities is important but challenging, especially as many tools have traditionally specialized in only a few data types. And while these complex datasets generate insights individually, integration with other –omics datasets is cruicial to help researchers discover and validate findings.

The UCSC Xena system was developed as a comprehensive high-performance visualization and analysis tool for both large public repositories and private datasets, built to scale with the current, and future, data growth and complexity. Xena's unique architecture enables cancer researchers of all computational backgrounds to explore large diverse datasets (Cieślik 2018, Langmead 2018) as well as their own data. Researchers use this same system to securely explore their own data, together or separately from the public data. The system easily supports many tens of thousands of samples and has been tested up to as many as a million cells. The simple and flexible architecture supports a variety of common and uncommon genomic and clinical data types. Xena hosts datasets from landmark cancer genomics resources including TCGA (The Cancer Genome Atlas) (Chin 2011, Chin 2011), ICGC (International Cancer Genome Consortium) (The International Cancer Genome Consortium 2010), and the GDC (Genomic Data Commons) (Grossman 2016). Xena's unique visualizations integrate gene-centric and genomic-coordinate-centric views across multiple data modalities, providing a deep, comprehensive view of genomic events within a cohort of tumors.

Results

UCSC Xena (http://xena.ucsc.edu) is a visual integration and exploration tool for public and private multi-omic data. The Xena platform has two components: the frontend web-based Xena Browser and the backend turn-key Xena Hubs (Figure 1). The Xena Browser empowers biologists to explore data across multiple Xena Hubs using a variety of visualizations and analyses. The back-end Xena Hubs host genomics data from laptops, public servers, behind a firewall, or in the cloud, and are configured to be public or private (Figure 1). The Xena Browser simultaneously connects to any number of Xena Hubs allowing data to be distributed across multiple Xena Hubs and visualized integratively. Integration occurs only within the browser, keeping private data secure.

This unique decoupled front-end Xena Browser and back-end Xena Hubs architecture has several advantages, especially over its predessor, the UCSC Cancer Browser. First, researchers can easily view their own private data by installing their own Xena Hub. Xena Hubs are lightweight compared to a full-fledged application and install easily on most computers, including a point-and-click installation interface for Windows and Mac computers. Second, users can use the same platform to view both public and private data together. Xena integrates data across multiple hubs, allowing users to view data from separate hubs as a coherent data resource. Xena does this while avoiding the download large public resources and keeping private data secure. This is especially useful for researchers who wish to view their own analysis results on public data, such as their own clustering or gene fusion calls, but don’t want to host a separate version of these large resources. Third, the Xena platform scales easily. As more datasets are generated, more Xena Hubs are added to the network, effectively growing with expanding genomics resources.

Turn-key Xena Hub

The backend Xena Hubs are designed to be turn-key, allowing users who may not be computationally savvy to easily install and use a Xena Hub on their personal computer, empowering them to view their own data (https://xena.ucsc.edu/private-hubs/). Xena Hubs run on most operating systems, including Windows, Mac and Linux. An interactive setup wizard guides users through the proces of installing and running a Xena Hub on their Windows or Mac computers, while a web wizard guides them through the data loading process, including collecting metadata relevant for visualization (https://tinyurl.com/localXenaHub). Users view and analyze their data using the same Xena Browser as those for public data. Hubs running on a personal computer, such as a laptop, are by default private, only allowing connections from the users' own Xena Browser, keeping data secure.

In addition to these user-friendly wizards, Xena Hubs can be installed and used via the command line. A dockerized version of the Xena Hub can also be used as part of an automated workflow pipeline to visualize computational results. Xena Hubs started in the cloud or on a server can be kept private by using a firewall. This enables easy sharing of private data within a lab, institution, or as part of a larger collaboration. They also can be configured to be public, making the data accessible to the larger community.

An example of a public hub hosted by an institution is the Treehouse Hub, which is deployed by the Treehouse project (https://treehousegenomics.soe.ucsc.edu). It hosts RNAseq gene expression data of pediatric cancer samples provided by Treehouse's clinical partners and repositories, harmonized with the UCSC RNAseq recompute compendium dataset of TCGA, TARGET and GTEx samples (Morozova 2017, Vivian 2017). This data is used to facilitate interpretation of a pediatric sample in the context of a large pan-cancer cohort, as all samples were processed by the same bioinformatics pipeline. Since the data hub is set up and hosted separately,

Treehouse has complete control over the data and data access. They use the UCSC Xena platform to host and update their public data, allowing it to be downloaded and visualized.

Performance is critical for interactive visualization tools, especially on the web. Increasing sample sizes for genomic experiments has become a challenge for many tools, including for the UCSC Cancer Browser. Knowing this, we optimized Xena Hubs to support data queries on many tens of thousands of samples, delivering slices of data in milliseconds to a few seconds. Tested using the TCGA Breast Cancer data, public hubs we deployed in the cloud have an average response rate of 244 ms with 50 users making concurrent requests.

Xena Browser

The Xena Browser (https://xenabrowser.net) is the online visualization and exploration tool for all Xena Hubs, both public and private. Its visualizations and analyses include the Xena Visual Spreadsheet, survival analysis, scatter plots, bar graphs, statistical tests, and genomic signatures. Researchers can dynamically generate and compare subgroups using Xena's sophisticated filtering and searching. Its shareable bookmarks and high resolution pdfs promote results dissemination and collaboration. In addition to Xena Browser's own views, we connect with a variety of complementary visualization tools, including the UCSC Genome Browser. We support modern web browsers such as Chrome, Firefox or Safari.

Visual Spreadsheet

It is essential to view different types of data, such as gene expression, copy number, and mutations, on genes or genomic regions, side-by-side. Integration across these diverse data modalities provides researchers a more biologically complete understanding of genomic events or

tumor biology. We designed our unique primary visualization, the Xena Visual Spreadsheet, to enable and enhance on this integration. Analogous to an office spreadsheet application, it is a visual representation of a data grid where each column is a slice of genomic or phenotypic data (e.g. gene expression or age), and each row is a single entity (e.g. a bulk tumor sample, cell line, or single cell) (Figure 2). Rows of these entities are sorted according to the genomics data being displayed. Researchers can easily re-order the Visual Spreadsheet as well as zoom in to just a few samples or out to the whole cohort, leading to an infinite variety of views in real time. These dynamic views enable the discovery of patterns among genomic and phenotype parameters, even when the data are hosted across multiple data hubs (Figure 2).

Xena's Visual Spreadsheet displays genomic data in gene-centric, coordinate-centric and feature-centric views. Gene-centric views show data mapped to a gene, such as copy number variation segments, gene expression, exons expression or mutations, and can display exonic regions only or include data mapped to introns (Figure 2, Column D, E, F, and G). Coordinate-centric views show the same data mapped along the genomic coordinate (Figure 2, Column C) and support genomic intervals from the base level up to an entire chromosome or chromosome arm. Feature-centric views show data on specific identifiers, such as ATACseq peaks or CpG probes (Supplemental Figure 1, column C). Feature-centric views can contain one or many features, as the user specifies, and can be hierarchally clustered. Gene-, coordinate-, and feature-centric views all support visualization of coding and non-coding regions (Supplemental Figure 2), as well as dynamic zooming to a region of interest. Links to the UCSC Genome Browser give genomic context to any gene, chromosome region, or feature. In addition to these genomics views, we also visualize phenotype and clinical data such as age, gender, expression signatures, cell types, and subtype classifications (Figure 2, Column B). These phenotypic data are crucial to enable users to go beyond

genomic-only discoveries. All columns and views can be placed side-by-side in a single Xena Visual Spreadsheet integrating diverse data over a cohort of samples.

The power of the Visual Spreadsheet is its deep data integration, which is not seen in other tools, including the UCSC Cancer Browser. Integration across different data modalities, such as gene expression, copy number variation, and DNA methylation, gives users a more comprehensive view of a genomic event in a tumor sample. For example, Xena's Visual Spreadsheet can help elucidate if higher expression for a gene is driven by copy number amplification or by a missense mutation (Supplemental Figure 3). Integration across gene- and coordinate-centric views helps users examine genomic events in different chromosome contexts. For example, Xena's Visual Spreadsheet can help elucidate if a gene amplification is part of chromosomal arm duplication or a focal amplification (Supplemental Figure 4). Integration across genomic and clinical data gives users the ability to make connections between genomic patterns and clinically relevant phenotypes such as subtypes. For example, Xena's Visual Spreadsheet can help elucidate if increased HRD signature scores are enriched in a specific cancer type or subtypes (Supplemental Figure 5). Finally, integration across user's own data and public resources on the same samples. For example, Xena's Visual Spreadsheet can help a researcher see how a fusion call from the literature relates to the expression of other downstream genes (Figure 2). These diverse integrations help researchers harness the power of comprehensive genomics studies, either their own or of public resources, driving discovery and a deeper understanding of cancer biology.

More browser visualizations and functionalities

In addition to the Visual Spreadsheet, Xena has many additional powerful views and analyses. Its Kaplan-Meier analysis allows users to statistically assess survival by any genomic or phenotypic data. Users can refine their analysis using a versatile set of options including selecting different

survival endpoints, multiple options to stratify continuous value parameters such as gene expression, and specify custom time to event cutoff such as 5 years or 10 years (Figure 3). The same data in a Xena Spreadsheet can also be viewed using bar charts, box plots and scatter plots, all with statistical tests automatically computed (chi-squared, t-test, or ANOVA, as appropriate), provide additional insights into the data (Supplemental Figure 5).

Our powerful text-based search allows users to dynamically highlight, filter, and group samples, a new feature not present in the UCSC Cancer Browser (Figure 4). Researchers can search the data on the screen similar to the ‘find’ functionality in Microsoft Word. Samples are matched and highlighted in real-time as the user types. Researchers can then filter, focusing the visualization to their samples of interest, or dynamically build a new two-group column, where samples are marked as 'true' or 'false', depending if they meet the researcher's search criteria. The new two-group column behaves like any other column and can be used in a Kaplan-Meier analysis, box plot, or other statistical analyses to compare the two groups of samples. This is a powerful way to dynamically construct two sub-populations based on any genomic data for comparison and analysis.

Being able to share and distribute biological insights is crucial, especially in this era of collaborative genomics. Xena’s bookmark functionality enables the sharing of live views. With a single click, users can generate a URL of their current view, which will take researchers back to the live browser session. We store the data in view for each URL so that it can be shared with colleagues or included in presentations. If a view contains data from a non-public Xena Hub, we allow users to instead download the current visualization state as a file. By giving users a file instead of a URL, we ensure that we never keep user’s private data on our public servers. This file can then be appropriately shared, and imported into the Xena Browser to recreate the live view. In addition to

bookmarks, researchers can generate a high resolution PDF figure of their current visualization for reports and publications.

Gene-expression signatures are developed by researchers to differentiate distinct subtypes of tumors, identify important cellular responses to their environment, and predict clinical outcomes in cancer (Sotiriou 2009). Xena's genomic signature functionality allows users to enter these signatures as a weighted sum of a marker gene set, a form commonly seen in publications, and dynamically build a new spreadsheet column of the resulting scores. This functionality allows researchers to test existing signatures or build new ones, allowing the comparison of a signature score with other genomic and phenotypic data.

The Transcript View enables comparison of transcript-level expression between two groups of samples, such as TCGA 'tumor' vs. GTEx 'normal', for all the transcripts of a gene (Supplemental Figure 6). We also provide context-dependent links to complementary visualizations such as the Tumor Map (https://tumormap.ucsc.edu) (Newton 2017) and MuPIT in CRAVAT (http://mupit.icm.jhu.edu/) (Niknafs 2013), enabling users to easily see a genomic pattern from a different perspective.

To assist researchers in building a Visual Spreadsheet, we developed a three-step guided wizard. This ensures that even new users who are unfamiliar with Xena can build basic visualizations. We also provide links to live examples that showcase useful and scientifically interesting visualizations, highlighting the power of Xena. We support our users by developing training videos, online and in-person workshops, and through help documentation. We keep users up-to-date on new features and datasets through social media, mailing lists and monthly newsletter.

Public Xena Hubs and supported data types

Today, cancer genomics research studies commonly collect data on somatic mutations, copy number, and gene expression, with other data types being relatively rare. However, as genomics technology advances, we expect these rarer data types to increase in frequency and new data types to be produced. With this in mind we designed Xena to be able to load any tabular or matrix formatted data, giving us exceptional flexibility in the types of data we can visualize, such as structural variant data (Supplemental Figure 7) and ATACseq peak signals (Supplemental Figure 1), a significant advantage over the UCSC Cancer Browser. Current supported data modalities include somatic and germline SNPs, INDELs, large structural variants, copy number variation, gene-, transcript-, exon-, protein-expression, DNA methylation, ATAC-seq peak signals, phenotypes, clinical data, and sample annotations.

UCSC Xena provides interactive online visualization of seminal cancer genomics datasets. To showcase these data resources, we deploy multiple public Xena Hubs. Together, they host over 1500 datasets from more than 50 cancer types, including the latest from TCGA (Hoadley 2018), ICGC, PCAWG (Pan-Cancer Analysis of Whole Genomes) (Campbell 2017), and the GDC (Table 1). Our TCGA Hub hosts data from TCGA, the most comprehensive cancer genomics dataset to-date, with a full set of data modalities for 12,000+ samples across 30+ cancer types. The Xena TCGA Hub hosts all public-tier TCGA derived datasets including somatic mutation, copy number variation, gene and exon expression, and more. Our Pan-cancer Altas Hub hosts data from the latest TCGA project, the Pan-Cancer Atlas, which conducted an integrative molecular analysis of the all tumors in TCGA. In addition to batch-effect-adjusted genomics datasets, the hub also hosts highly curated datasets such as molecular subtypes and multiple survival endpoints (Hoadley 2018, Ding 2018, Sanchez-Vega 2018). Our ICGC Hub hosts data from the ICGC project, a global effort to create a comprehensive

description of the genomic, transcriptomic and epigenomic changes in 50 different tumor types. Our PCAWG Hub supports PCAWG, an analysis of 2,600 ICGC whole-cancer genomes and their matching normal tissues across 39 distinct tumour types. Its datasets include somatic mutation data from the whole genome, large structural variants, RNAseq-based data analysis, mutational signatures, curated histology, and more. Our GDC Hub hosts data from GDC, a uniformly recomputed dataset using uniform bioinformatics pipelines and the latest human genome assembly, hg38. In addition to these well-known resources, we also host results from the UCSC Toil RNAseq recompute compendium, a uniformly re-aligned and re-called gene and transcript expression dataset for all TCGA, TARGET and GTEx samples (Vivian 2017). This dataset allows users to compare gene and transcript expression of TCGA 'tumor' samples to corresponding GTEx 'normal' samples. Lastly, the UCSC Public Hub has data curated from various publications such as CCLE (Cancer Cell Line Encyclopedia, Barretina 2012) and MET500 (Robinson 2017). Xena Hubs load only the derived datasets, leaving the raw sequencing data at their respective locations. Xena complements each of these resources by providing powerful interactive visualizations for these data. All public Xena Hubs (https://xenabrowser.net/hub/) are open access, with no account or login required.

In addition to visualization, these public data hubs support data download in bulk for downstream analyses. We also offer programmatic access to slices of data through the Xena python package (https://github.com/ucscXena/xenaPython), which can be used independently or in a Jupyter Notebook to access any of the public Xena Hubs.

Discussion

UCSC Xena is a tool for cancer researchers to explore, visualize, and analyze functional genomics data, whether it is a public resource or their own data. The Visual Spreadsheet,

sophisticated filtering and subgrouping, and Kaplan-Meier analysis enable researchers of all computational backgrounds to investigate complex genomics datasets. We support virtually all data modalities including mutations, copy number, expression, phenotype and clinical data as well as rare data types such as non-coding mutations, large structural variants, DNA methylation, and ATAC-seq peak signals. Integration across different data modalities and visualizations, as well as between genomic and clinical data yield insightful views into cancer biology. We host many large public datasets, such as TCGA, Pan-Cancer Atlas, PCAWG, GDC, and ICGC, helping to make these powerful resources accessible to investigators. UCSC Xena departs from its predecesor, the UCSC Cancer Browser through its unique architecture, its support for datasets with more than ~1,000 samples, and its views of unusual and diverse data types. Xena also has a number of unique visualizations and analyses including the Visual Spreadsheet, powerful filtering and subgrouping, and the Transcript View.

UCSC Xena complements existing tools including the cBio Portal (http://www.cbioportal.org/, Cerami 2012), ICGC Portal (https://dcc.icgc.org/, Zhang 2011), GDC Portal (https://portal.gdc.cancer.gov/, Jensen 2017), IGV (http://software.broadinstitute.org/software/igv/, Thorvaldsdóttir 2013), and St. Jude Cloud (https://stjude.cloud/, Ma 2018) in a number of ways. First, our focus on providing researchers a lightweight, easy-to-install platform to visualize their own data as well as data from the public sphere. By visualizing data across multiple hubs simultaneously, Xena differentiaties itself from other tools by enabling researchers to view their own data together with consortium data while still maintaining privacy. Further, Xena focuses on integrative visualization of multi-omics datasets across different genomic contexts, including gene, genomic element, or any genomic region for both coding and non-coding part of the genome. Finally, Xena is built for performance. It can easily visualize of tens of thousands of samples in a few seconds and has been

tested on single-cell data with up to a million cells. With single-cell technology, datasets will become orders of magnitude larger than traditional bulk tumor samples. Xena is well-positioned to rise to the challenge and we anticipate further performance optimizations to support web-based data visualization of million-cell datasets.

While it is widely recognized that data sharing is key to advancing cancer research, how it is shared can impact the ease of data access. UCSC Xena is a designed for cancer researchers both with and without computational expertise to easily share and access data. Users without a strong computational background can explore their own data by installing a Xena Hub on their personal computer using our installation and data upload wizards. Bioinformaticians can install a private or public Xena Hub on a server or in the cloud or as part of an analysis pipeline, making the generated data available in a user-friendly manner that requires little extra effort. Data sharing has, and will continue to, advance cancer biology and Xena is part of the technological ecosystem that supports this.

UCSC Xena is a scalable solution to the rapidly expanding and decentralized cancer genomics data. Xena's architecture, with it’s detached data hubs and web-browser-based visualization, allows new projects to easily add their data to the growing compendium that we support. Additionally, by maintaining a flexible input formats, we support many different data modalities, both now and in future. Xena excels at viewing cohorts of samples, cells, or cell lines and showing trends across those entities, whether they be human or a model organism. While we have focused on cancer genomics, the platform is general enough to host any functional genomics data. In this age of expanding data resources, Xena's design supports the ongoing needs of the cancer research community.

Methods

Xena Hub

The Xena Hub is a JVM-based application, written in Clojure, that serves functional genomic data over HTTP. It exposes a relational query API for data slicing and metadata. We use a query language instead of REST for our APIs because it allowed us to decouple the client and server. To support interactive visualization, REST APIs would have to be denormalized for performance (e.g. by joining related objects, and projecting the result). This would require a tight coupling between the REST endpoints and particular views, and therefore frequent Xena Hub updates to match browser client updates. This was untenable given the numbers of Xena Hubs in public and private use. By using a query language we are able fetch exactly the data we need, and only the data we need, for quickly evolving visualizations and data shapes, without redeployment of the hubs. This is similar in motivation to Facebook's GraphQL, and Netflix's Falcor, but predates them.

Internally, Xena Hubs use the H2 database for storage. Data is stored in opaque blocks in a column orientation, which allows fast retrieval of a field for all samples of a dataset, or a subset of samples. A Hub can be installed either via the command line or via the point-and-click install4j graphical user interface wizard. A wizard process guides users through the data loading process for Hubs on a user’s own personal computer (https://xena.ucsc.edu/private-hubs/).

Xena Browser

The Xena Browser is a javascript web-browser application to visualize and analyze functional genomics data stored in one or more Xena Hubs. The primary technologies are React, the 2D canvas

API, and RxJS. Babel is used for es6 support, and webpack for the build. The application architecture is an asynchronous model similar to redux-observable (https://redux-observable.js.org/), with semantic actions that update application state, and action side-effects creating Rx streams that will dispatch later actions. The redux (https://redux.js.org/), or Om (https://github.com/omcljs/om), pattern of immutable, single-atom state makes it simple to keep multiple views in sync, and provides “time travel” debugging during development.

We use the canvas API because it performs better at large data scales than SVG libraries such as D3. With the advances in javascript JIT compilers, we find that optimized loops over canvas pixel buffers out-perform geometric drawing primitives, such as *rect()*, and *stroke()*, when rendering dense views of large data.

The jsverify property-based testing library (http://jsverify.github.io/) is used for unit and integration testing. Property-based, or "generative" testing is similar to fuzzing -- generating random test cases, and asserting invariants over the results -- but on failure, attempts to find a minimal failing test case. This usually results in more tractable failure cases. Property-based testing allows us to test a much larger portion of the input space than conventional "known-answer" unit tests, and frequently identifies failure cases that we would never think to test.

We have contributed two of our javascript modules to BioJS (Gómez 2013), including a Kaplan-Meier module (https://github.com/ucscXena/kaplan-meier) to compute Kaplan-Meier statistics, and a static-interval-tree library (https://github.com/ucscXena/static-interval-tree) to effectively find overlapping intervals in one dimension.

Code availability

The entire Xena codebase is open source and available for reuse under and Apache 2.0 license at https://github.com/ucscXena. Xena Hub code is available at https://github.com/ucscXena/ucsc-xena-server and Xena Browser code is available at https://github.com/ucscXena/ucsc-xena-client.

User-centered design principles

The Xena System was developed using User Experience Design methodologies. User-Centered Design is design based upon an explicit understanding of users, tasks, and environments, and is driven and refined by iterative user-centered evaluation. We use need-finding interviews, prototypes, wireframes, and user acceptance testing to help ensure that Xena meets user needs.

Public Xena Hubs

We download functional genomics source data from each respective source: GDC data portal (https://portal.gdc.cancer.gov/repository) for the GDC Hub, GDC legacy archive (https://portal.gdc.cancer.gov/legacy-archive) for the TCGA Hub, ICGC data portal (https://dcc.icgc.org/) for the ICGC Hub, GDC PanCanAtlas Publications web site (https://gdc.cancer.gov/node/905/) for the Pan-Cancer Atlas Hub, Synapse ICGC-TCGA Whole Genome Pan-Cancer Analysis project (https://www.synapse.org/#!Synapse:syn2351328/wiki/62351) for PCAWG hub, and various publications for data hosted on the UCSC Public Hub and UCSC Toil RNAseq Recompute Hub. The GDC and ICGC Hubs are updated periodically.

The downloaded data were wrangled into a generic tabular or matrix format, and loaded into the corresponding Xena Hubs. Specific wrangling steps, including any normalization, is listed for each

dataset in the Xena Browser dataset pages (https://xenabrowser.net/datapages/). The wrangled data is available for bulk download from the dataset pages.

We deploy all public-facing Xena Hubs supported by our team in the Amazon Web Services (AWS) cloud-computing environment. Each hub is built using an AWS elastic load balancer connected to two EC2 r4.4xlarge servers with a single database stored on AWS elastic file system.

Acknowledgements

Research reported in this publication was supported by National Cancer Institute of the National Institutes of Health under award numbers 5U24CA180951-04 and 5U24CA210974-02. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This project has also been made possible in part by grant number 2018-182812 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation. We would also like to thank AWS Cloud Credits for Research and Google Summer of Code.

References

Barretina, J., Caponigro, G., Stransky, N., Venkatesan. K., Margolin. A.A., et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603-607 (2012).

The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511,** 543–550 (2014).

Campbell, P. J., Getz, G., Stuart, J. M., Korbel, J. O., Stein, L. D., et al. Pan-cancer analysis of whole genomes. Preprint at https://www.biorxiv.org/content/early/2017/07/12/162784 (2017).

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E, Sumer, S. O., et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery* **5,** 401-404 (2012).

Chin, L., Hahn, W.C., Getz, G. & Meyerson, M. Making sense of cancer genomic data. *Genes & Development* **25,** 534-555 (2011).

Chin, L., Andersen, J.N. & Futreal, P.A. Cancer genomics: from discovery science to personalized medicine, *Nature Medicine* **17,** 297-303 (2011).

Cieślik, M. & Chinnaiyan, A. M. Cancer transcriptome profiling at the juncture of clinical translation. *Nature Reviews Genetics* **19,** 93–109 (2018).

Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., e al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, 6413 (2018)

Ding L., Bailey, M. H., Porta-Pardo, Eduard, Wheeler, D. A., Getz, G. et. al. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* **173,** 305-320 (2018)

Gao, Q., Liang, W.W., Foltz, S.M. et al. Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep.* **23,** 227-238.e3 (2018)

Gómez, J., García, L. J., Salazar, G. A., Gore, J. V. S., García, A., et al. BioJS: an open source JavaScript framework for biological data visualization. *Bioinformatics* **29,** 1103–1104 (2013)

Grossman, R.L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., et al. Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine* **375,** 1109-1112 (2016).

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173,** 291–304 (2018).

The International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **464,** 993–998 (2010).

Jensen, M. A., Ferretti, V., Grossman, R. L. & Staudt, L. M. The NCI Genomic Data Commons as an engine for precision medicine. *Blood* **130,** 453-459 (2017).

Kong-Beltran, M., Seshagiri, S., Zha, J., Zhu, W., Bhawe, K., et al. Somatic mutations lead to an oncogenic deletion of met in lung cancer. *Cancer Research* **66,** 283-289 (2006).

Langmead, B. & Nellore, A. Cloud computing for genomic data analysis and collaboration. *Nature Reviews Genetics* **19,** 208–219 (2018).

Ma, X., Liu, Y., Liu, Y., Alexandrov, L. B., Edmonson, M.N., et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555,** 371-376 (2018).

Malta, T. M., Sokolov, A., Gentles, A. J., et al., Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell* **173,** 338–354 (2018).

Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24,** 133-141 (2008).

Mathelier, A., Lefebvre, C., Zhang, A. W., Arenillas, D. J., Ding, J., et al. Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. *Genome Biology* **16,** 84 (2015).

Morozova, O., Salama, S. R., Bjork, I., Goldstein, T. C., Mueller, S., et al. Comparative genomic analysis for pediatric cancer patients evaluated in a California Initiative to Advance Precision Medicine Demonstration Project. *Journal of Clinical Oncology* **35,** TPS10578-TPS10578 (2017)

Newton, Y., Novak, A. M., Swatloski, T., McColl, D. C., Chopra, S., et al. TumorMap: Exploring the Molecular Similarities of Cancer Samples in an Interactive Portal. *Cancer Research* **77,** e111–114 (2017)

Niknafs, N., Kim, D., Kim, R., Diekhans, M., Ryan, M., et al. MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures. *Human Genetics* **132,** 1235–1243 (2013)

Robinson, D.R., Wu, Y.M., Lonigro, R.J., Vats, P., Cobain, E., et al. Integrative clinical genomics of metastatic cancer. *Nature* **548,** 297-303 (2017).

Sanchez-Vega, F., Mina, M., Armenia, J., Ciriello, G., Sander, C., Schultz, N. et. al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173,** 321-337 (2018).

Schroeder, M. P., Gonzalez-Perez, A. & Lopez-Bigas, N. Visualizing multidimensional cancer genomics data. *Genome Medicine* **5,** 9 (2013).

Sotiriou, C. & Pusztai, L. Gene-Expression Signatures in Breast Cancer. *New England Journal of Medicine* **360,** 790-800 (2009).

Stephens, Z. D., Lee, S. L., Faghri, F., Campbell, R. H., Zhai, C., et al. Big Data: Astronomical or Genomical? *PLOS Biology* (2015).

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14,** 178-192 (2013).

Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C., Armstrong, J., et al. Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology* **35,** 314-316 (2017).

Wills, Q. F & Mead A. J. Application of single-cell genomics in cancer: promise and challenges. *Human molecular genetics* **24,** R74-R84 (2015).

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., et al. International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database* **2011,** (2011).

Zhou, W., Dinh, H.Q., Ramjan, Z., Weisenberger, D.J., Nicolet, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nature Genetics* **50**, 591-602 (2018).

Figure 1

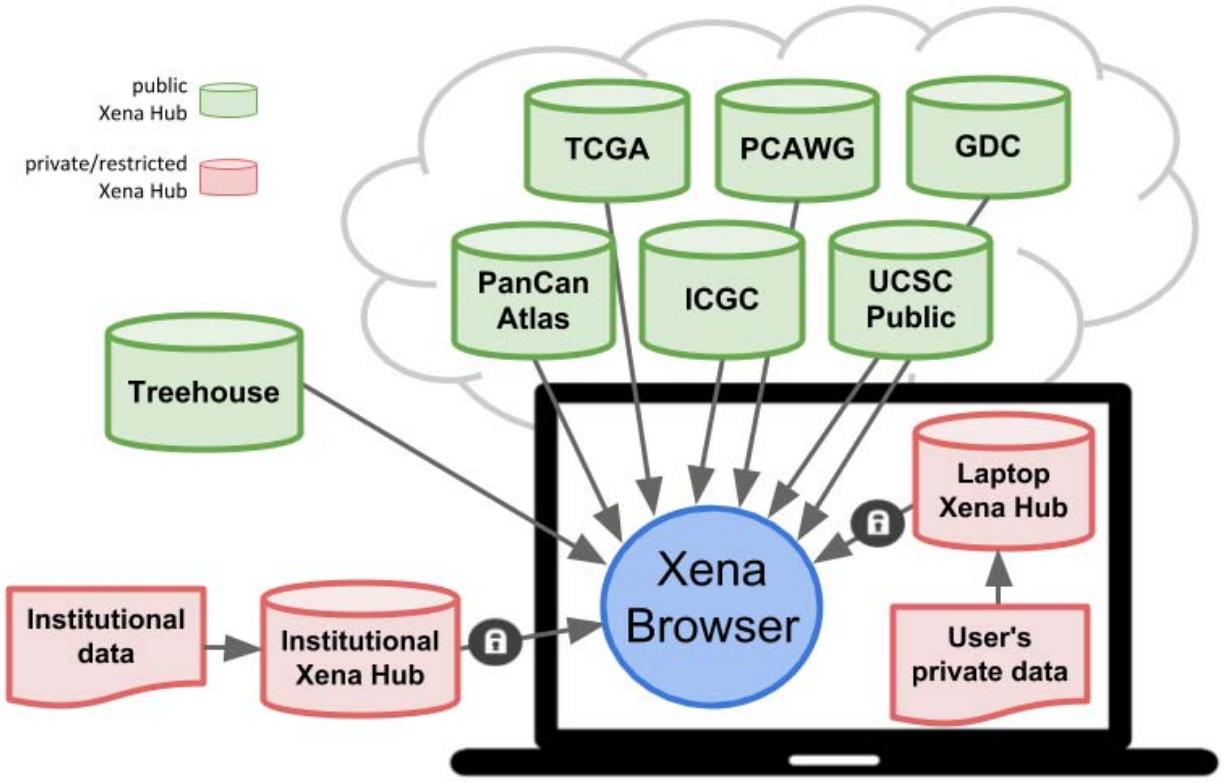


Figure 1. Diagram of the UCSC Xena platform architecture. Multiple Xena Hubs (each shown as a database icon) are connected to the Xena Browser simultaneously. Public hubs are in green and and private hubs in red. In this example, private data from an independent research collaboration (in red) is loaded into their own private Xena Hubs on their servers. Similarly, user's private data (in red) is loaded into a private Xena Hub on a researcher's computer. Data integration occurs within the Xena Browser on the user's computer. The lock icon indicates that only authorized users have access to the private Xena Hubs. Data only flows from hub to browser. This design achieves data integration across both public and private resources while maintaining each hub's data confidentiality.

Figure 2

Figure 2. An example Xena Browser Visual Spreadsheet examining published TCGA prostate cancer ERG-TMPRSS2 fusion calls by combining data from local and public Xena Hubs together. A user downloaded ERG-TMPRSS2 fusion calls from Gao et al. 2018 (n=492), loaded the data into their own local Xena Hub (column B), and then compared the fusion calls to public gene expression and copy number data from the same sample set (columns C-H). Columns C and D are the copy number variation status, first of a zoomed in region of chromosome 21 and second of just the ERG gene. The gene diagram at the top with exons in black boxes, tall coding regions, and shorter untranslated regions. Amplifications are in red and deletions are in blue. Columns E and F are gene expression, first for the ERG gene and second for each of the exons of the ERG gene. They are colored red to blue for high to low expression. Column G is SPOP mutation status and also has a gene diagram at the top. The position of each mutation is marked in relation to the gene diagram and colored by its

functional impact: deleterious mutations are red, missense and in-frame INDELs are blue, splice mutations are orange, and synonymous mutations are green. Here we can see that the fusion calls are highly consistent with the characteristic overexpression of ERG (columns E, F). However, only a subset of those samples in which a fusion was called can be seen to also have the fusion event observed in the copy number data via an intra-chromosomal deletion of chromosome 21 that fuses TMPRSS2 to ERG as shown in columns C and D. This observation is consistent with the 63.3% validation rate described in the paper. SPOP mutations (blue tick marks in column G) are mutually exclusive with the fusion event. A live Xena Browser view of the figure is at https://xenabrowser.net/?bookmark=fa40b4f1c016a7567e2a22ddbedbbb3b, created using the bookmark functionality. Rows are sorted by the left-most data column (column B) and subsorted on columns thereafter.

Figure 3

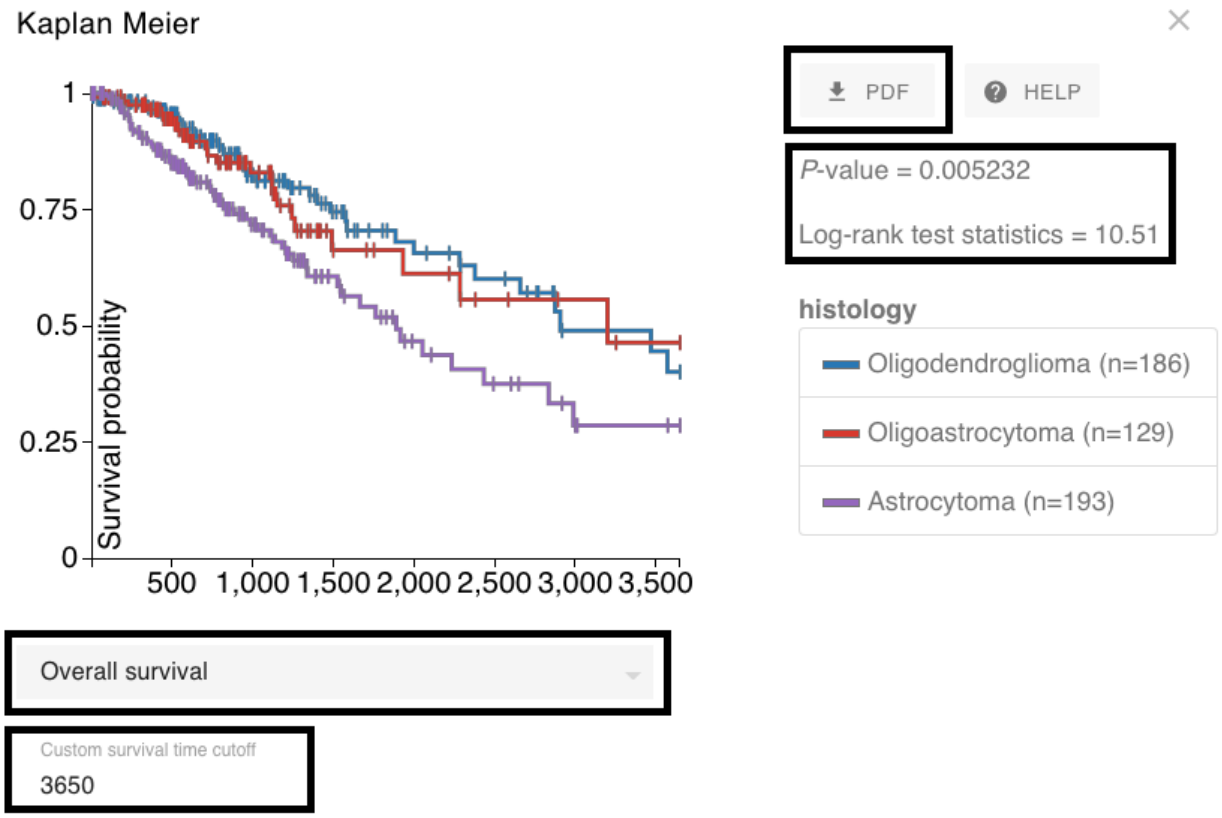
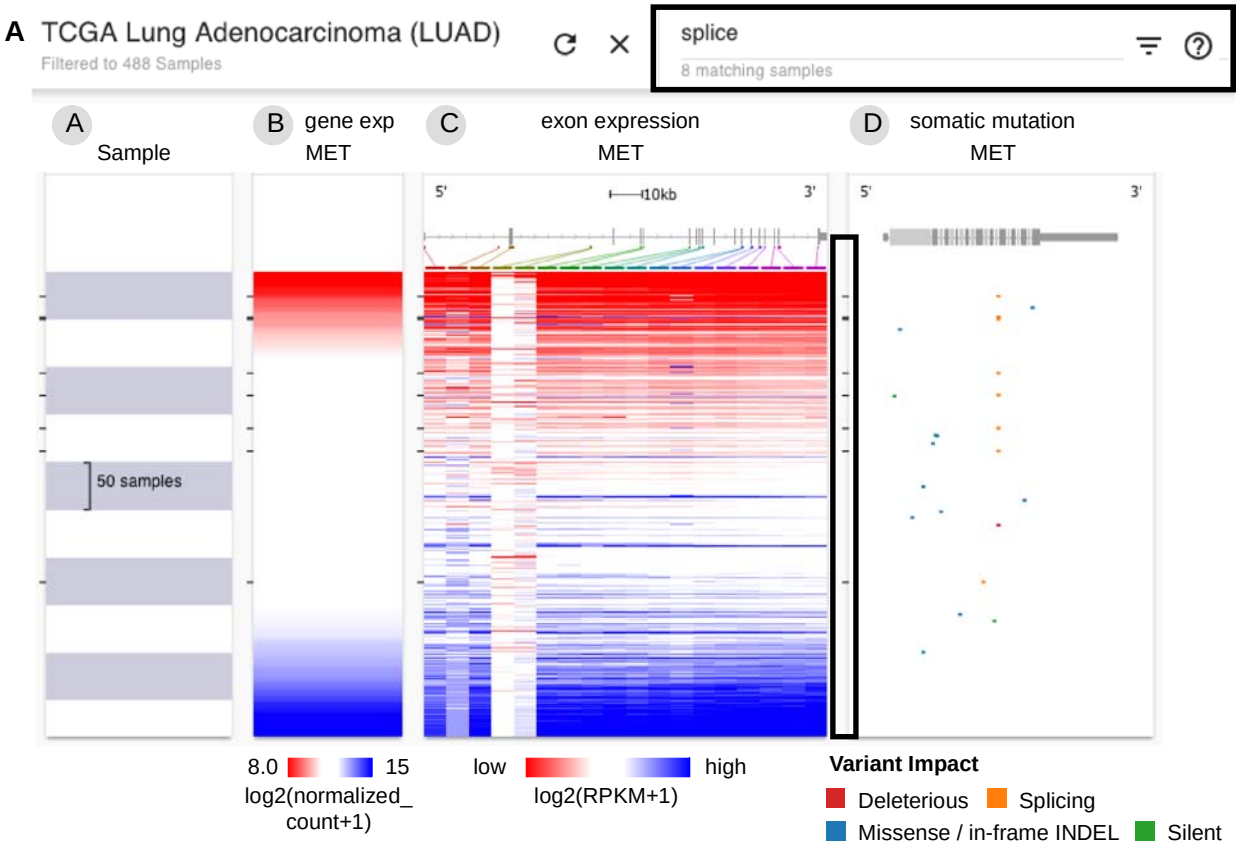


Figure 3. Kaplan-Meier analysis of overall survival for TCGA lower grade glioma histological subtypes. Black boxes in the figure highlight, top to bottom, a button to generate a PDF, the statistical analysis results, a dropdown menu to select different survival endpoints such as overall or recurrence-free survival, and a textbox to enter a custom survival time cutoff (currently set to 3,650 days, or 10 years). This figure shows that patients characterized as having the astrocytoma histological subtype have significantly worse 10-year overall survival compared to the oligodendroglioma and oligoastrocytoma subtypes ($p < 0.05$).

https://xenabrowser.net/heatmap/?bookmark=2f9d783982879594dd0f52564058372d

Figure 4



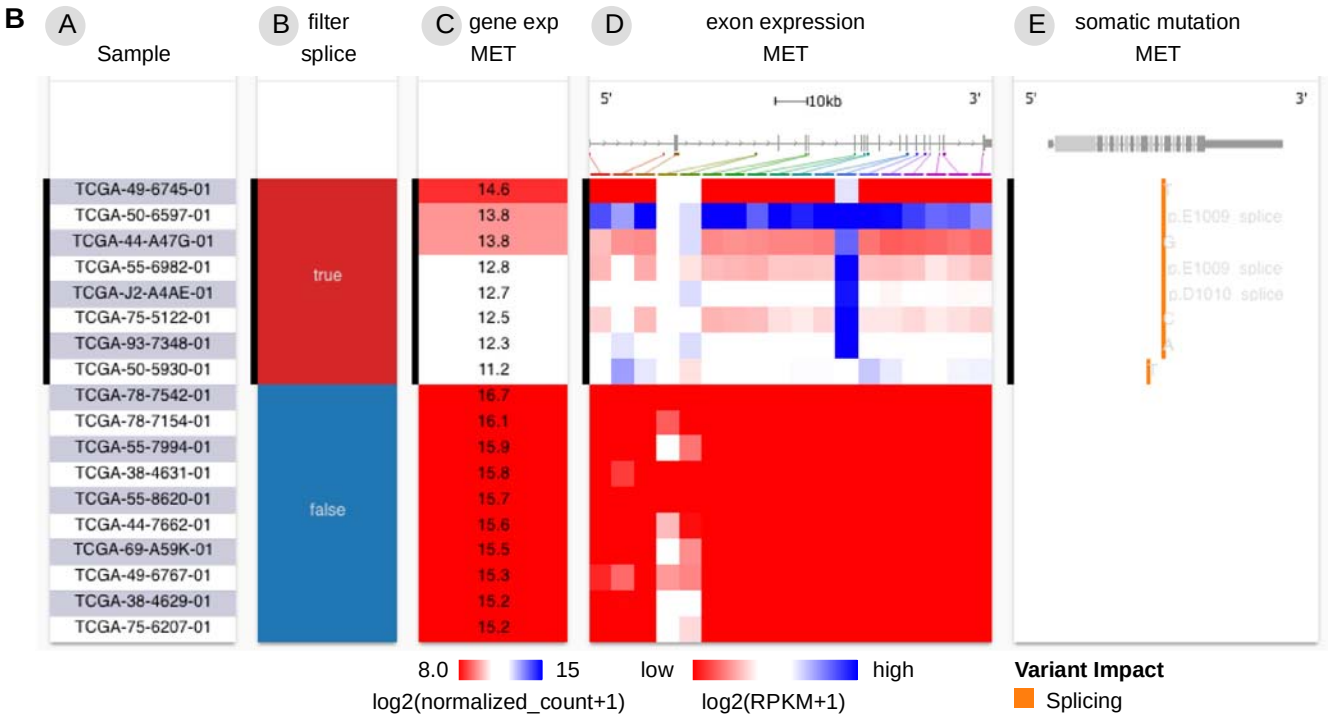


Figure 4. Xena Browser text-based find, highlight, filter, and subgroup samples functionality. **(A)** Finding and highlighting samples in TCGA lung adenocarcinoma cohort that have a splice mutation in the gene *MET*. Similar to the 'find in document' feature in Microsoft Word, users can search all data on the screen. In this figure, the Xena Browser searched all columns for the user's search term 'splice' and highlighted samples with a 'splice' mutation with black tick marks (highlighted by the black box). More complex search terms can include 'AND', 'OR', '>', '<', '=', and more. Users can dynamically filter, zoom, and create subgroups based on the search results. Columns from left to right are MET gene expression, MET exon expression and MET somatic mutation status.

https://xenabrowser.net/heatmap/?bookmark=c5873fb094ef714e44e65df217e93071

(B) After creating a new column with two subgroups. Columns are same as (a) with the user-generated column inserted on the left. Samples that matched the query of 'splice' were assigned a value of "true" and those that do not "false". The researcher has zoomed to a few samples at the top for a more detailed view. The figure shows that samples that have the splice site mutation (orange tick marks, column E) have lower expression of MET's exon (column D). The splice mutation causes exon 14 skipping and results in the activation of MET (Kong-Beltran 2006, The Cancer Genome Atlas Research Network 2014).

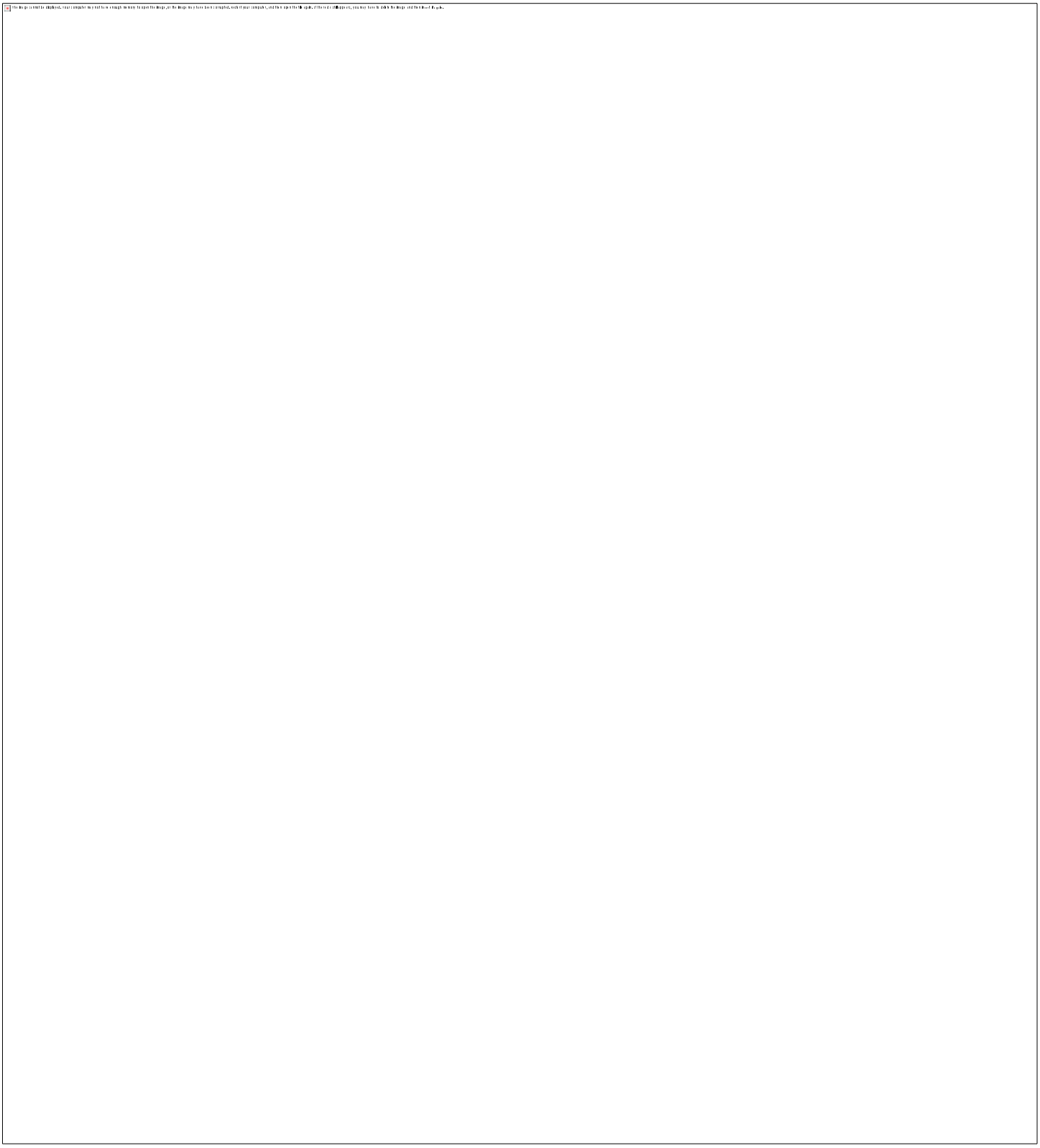
https://xenabrowser.net/heatmap/?bookmark=748c42b0b49552004da53873950aad62

Table 1

Public Xena Resources	Samples	Data Types
TCGA	12,470	copy number, gene-, exon-, miRNA-, and protein-expression, somatic mutation, DNA methylation, survival, and clinical data
Pan-Cancer Atlas	12,591	copy number, gene-, miRNA-, and protein-expression, somatic mutation, DNA methylation, molecular subtypes, curated survival data
ICGC	17,697	copy number, gene expression, somatic mutation
PCAWG	2,834	whole-genome copy number, somatic mutations, large structural variants, gene- and miRNA-expression, purity, ploidy, mutational signatures, survival, and curated histology
UCSC RNA-seq	19,131	TCGA, TARGET, and GTEx gene- and transcript-expression
GDC	20,157	TCGA and TARGET copy number, somatic mutations, gene- and miRNA-expression, DNA methylation, overall survival, and clinical data
TCGA ATAC-seq	404	ATAC-seq peak signal

Table 1. Summary of data hosted on Public Xena Hubs as of October 26, 2018.

Supplemental Figure 1

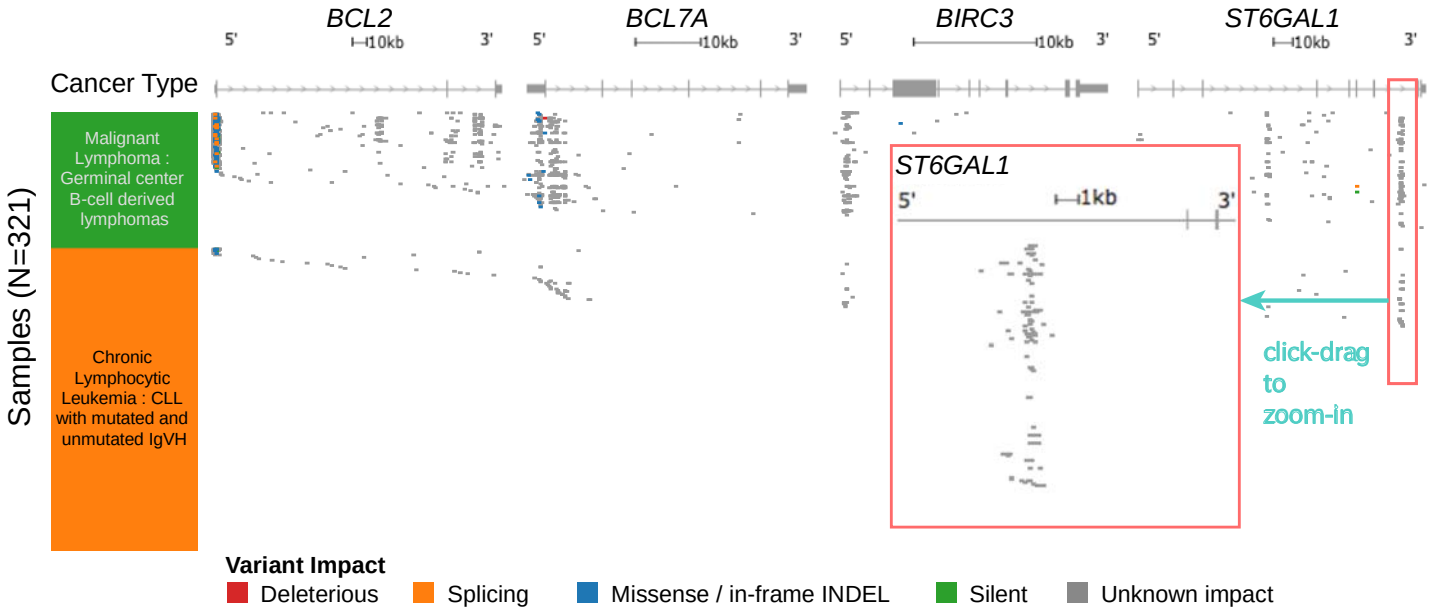


Supplemental Figure 1. (A) Xena Visual Spreadsheet examining the relationship between MYC gene expression (Column G), chromatin accessibility (ATAC-seq peaks in Column C through F, gene model at the top showing red to blue for increased to decreased chromatin accessibility), and cancer type (Column B). COAD and KIRC samples are highlighted with black bars across all columns using Xena's dynamic find and highlight feature. Here we see that rs6983267, a functionally validated GWAS colon and prostate cancer susceptibility SNP, is more accessible in COAD, especially when compared to KIRC (Column C). Looking at the surrounding peaks (Column D) we can see that this pattern is localized to this SNP. Accessibility at this SNP is also associated with accessibility at MYC (Column E) and specifically the MYC promoter (Column F). Looking across all cancer types we see these COAD-specific findings may be applicable to other cancer types with extensive chromatin accessibility at 5' and 3' DNA elements, as suggested in Corces et al 2018.

https://xenabrowser.net/heatmap/?bookmark=032e821bfa82eb3c0f522877dbe3b252

(B) Zooming in on COAD and KIRC we can more clearly see the contrast between these two cancer types. https://xenabrowser.net/heatmap/?bookmark=471618c9813dabdee556bbaae43cf883

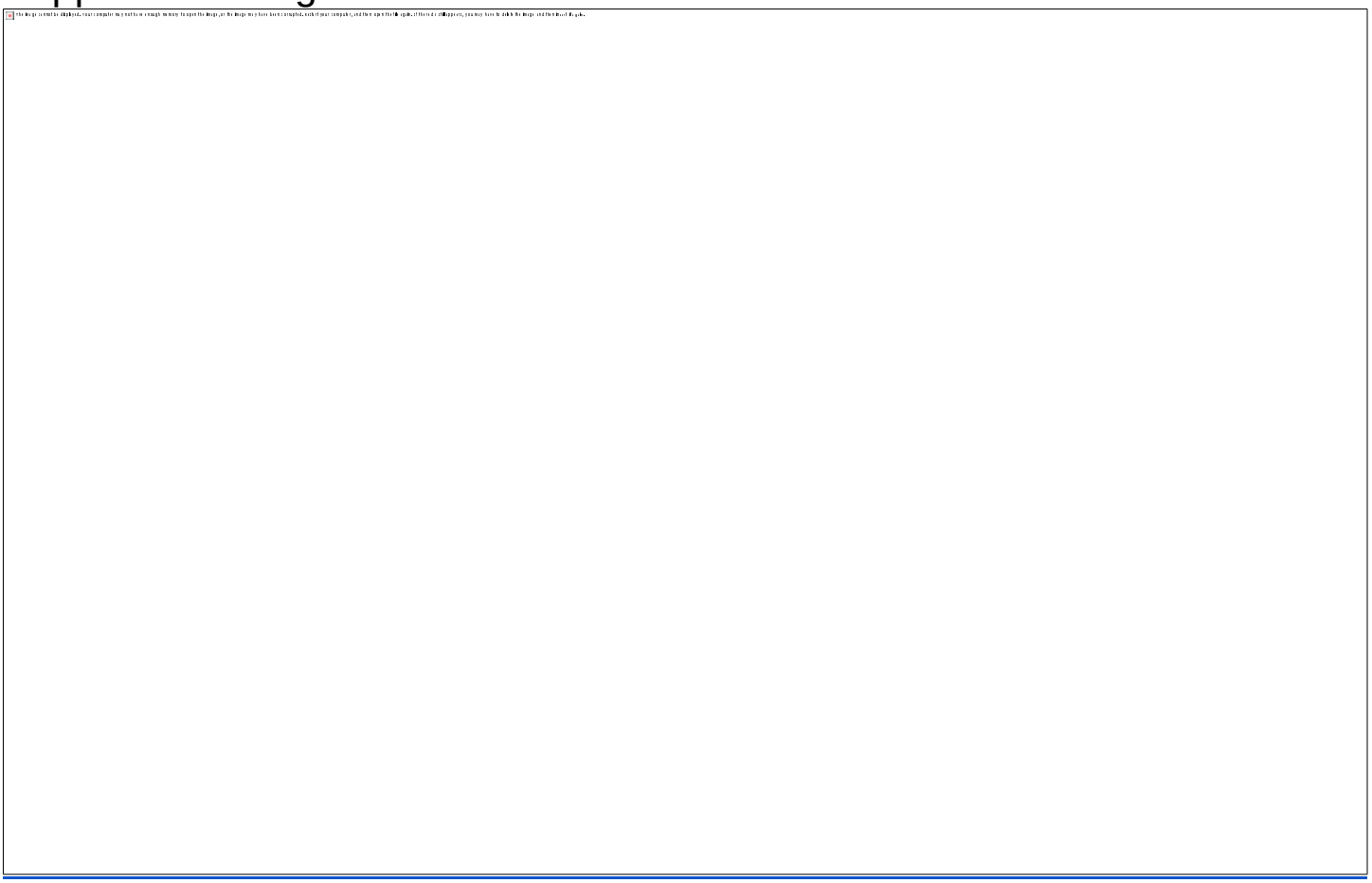
Supplemental Figure 2



Supplemental Figure 2. Visualization of both coding and non-coding mutations from a gene-centric perspective in ICGC lymphomas. The columns left to right are cancer type, BCL2, BCL7A, BIRC3 and ST6GAL1 mutation status, respectively. Gene diagrams are shown at the top of each column, with exons as gray boxes, introns as lines. The position of each mutation is marked in relation to the gene diagram and colored by its functional impact. This figure shows the intronic mutations hotspots in these genes. These mutation 'pile-ups' would be not be visible if viewing exomes only. A dynamic toggle allows user to show or hide introns from the view. While the majority of the intronic mutations in this view have an unknown impact (shown in grey), they overlap with known enhancers regions (Mathelier 2015). Insert is a zoomed-in view of one of the hotspots in ST6GAL1. Users can click-and-drag to trigger zoom.

https://xenabrowser.net/heatmap/?bookmark=a11909d2c2c629ee999e1a9802fac7dd

Supplemental Figure 3



Supplemental Figure 3. (A) Xena Visual Spreadsheet showing EGFR status in TCGA Lung Adenocarcinoma. Column B is the EGFR subgroup where blue are EGFR wild-type samples, red are samples with an EGFR missense mutation or in frame deletion, purple are samples that have an amplification of EGFR (copy number value greater than 0.5), and orange are samples that have both an amplication and altered mutation status. Column C is EGFR copy number variation status, Column D is EGFR mutation status, and Column E is EGFR expression. This view shows that high EGFR expression can be driven by copy number amplification, mutation status, or both.

https://xenabrowser.net/?bookmark=5f3af5136b4ee90871da0b7fb81cb083

(B) Here we can see the average EGFR expression for each of the 4 EGFR subgroups. Samples with both an amplification or altered mutation status have the highest gene expression.

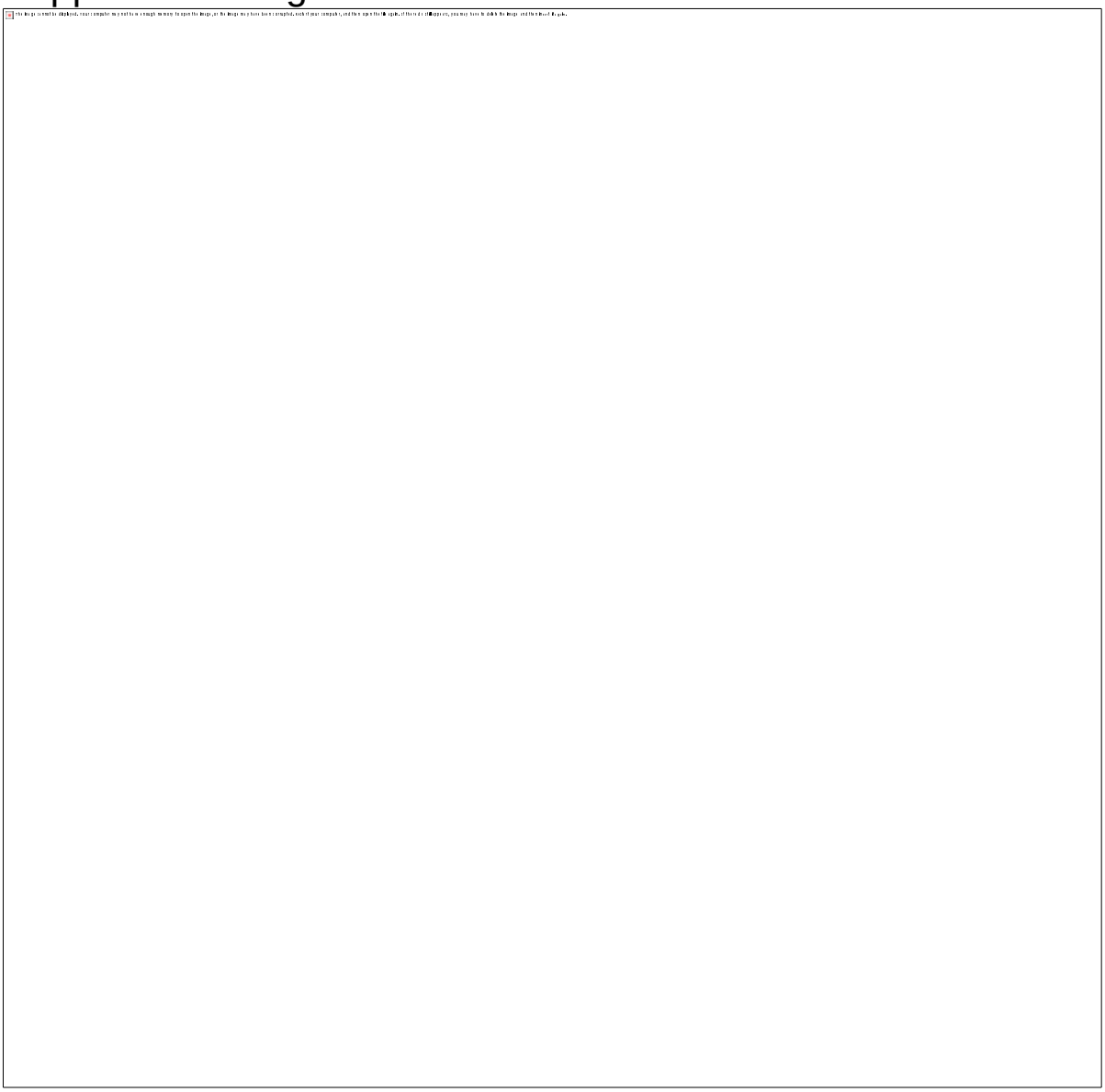
https://xenabrowser.net/?bookmark=45b31e1e4fa268412abb7040c81057fb

Supplemental Figure 4

Supplemental Figure 4. A Xena Visual Spreadsheet contrasting gene-level and chromosome-arm-level copy number amplifications in TCGA Breast Cancer. Columns B through D show copy number status for just ERBB2, the entirety of chromosome 8, and ERBB2 and its flanking regions. We can see in column B that some samples have amplifications in ERBB2, but that this amplification is local to the ERBB2 gene (comparing Column B to Column D). In contrast we can see that many samples have an amplification of the entirety of the p-arm of chromosome 8, which is not localized to any one gene. In the upper right we can see an example of Xena's tooltip, which shows more information about the data under the mouse cursor.

https://xenabrowser.net/?bookmark=9127f68a4b1547cad2bc493b97508416

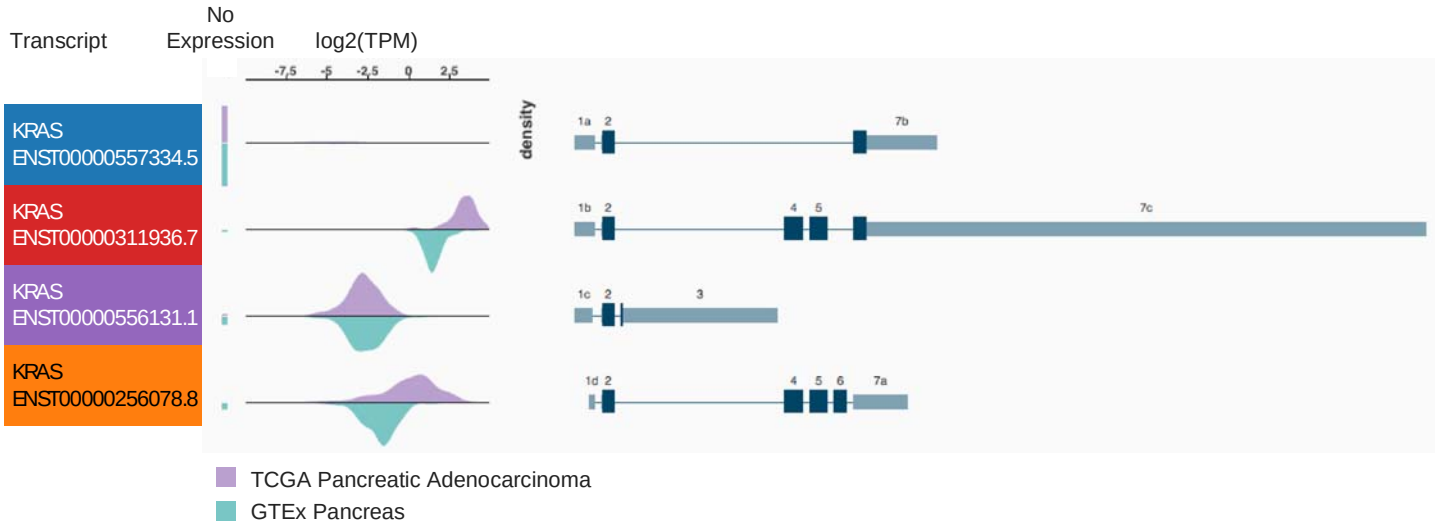
Supplemental Figure 5



Supplemental Figure 5. Stemness score (RNA based) as called by the PanCan Atlas Project across cancer type (Malta 2018). Cancers progress in part through the gradual gain of stem-cell-like features. Malta 2018 provides novel stemness indices for assessing this oncogenic de-differentiation. Here we can see that Testicular Germ Cell Tumors (TGCT) has a highest median stemness score compared to all other cancer types.

https://xenabrowser.net/?bookmark=7deeea053f221ba9fc506be855f353ac

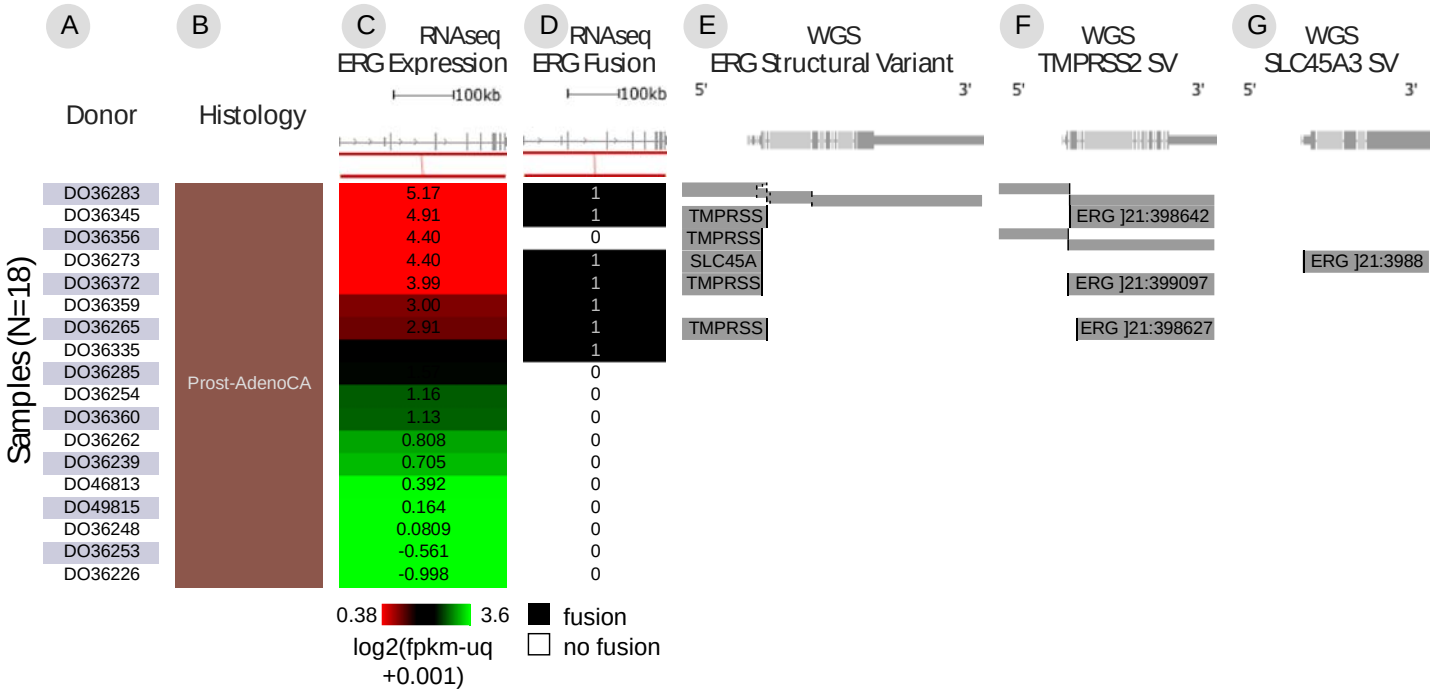
Supplemental Figure 6



Supplemental Figure 6. Transcript View in Xena showing four KRAS transcripts' expression for TCGA pancreatic adenocarcinoma and GTEx normal pancreas tissue. To generate a view, a researcher enters a gene and select two populations. The visualization will display, for all transcripts for that gene, a double (top and bottom) density distribution of transcript expression in each population. We see that for KRAS, transcript ENST00000311936.7 (second from the top), has higher expression in pancreatic tumors (TCGA) compared to the normal pancreas tissue (GTEx).

https://xenabrowser.net/transcripts/?bookmark=b832d9683cd8914aad0215b616bd5b21

Supplemental Figure 7



Supplemental Figure 7. Visualization of large structural variants. This figure shows frequent ERG fusion in PCAWG prostate cancer detected by both RNA-seq and DNA-seq analysis. The left three data columns (B, C, D) are histology, ERG gene expression, and ERG fusion detected using RNA-seq data. In the ERG fusion column (D), samples that have a fusion are marked with 1 and those that do not are marked with 0. The next three columns (E, F, G) show structural variant calls made using whole-genome DNA-seq data for ERG, TMPRSS2, and SLC45A3. Precise breakpoints are mapped to gene diagrams. A grey bar indicates an external piece of DNA that is fused at the breakpoint. Gene names on the grey bars show the origin of the external DNA that is joined. This figure shows that TMPRSS2 and SLC45A3 are fusion partners for ERG, and that these fusions correlate with over-expression of ERG. Fusions detected by RNA-seq and whole-genome sequencing are not always consistent. Here, even using a consensus of DNA-based detection methods, one fusion detected by a consensus of RNA-based detectors is missed, and the converse is also seen. This example shows

that an integrated visualization across multiple data types and algorithms provides a more accurate model of a genomic event.

https://xenabrowser.net/heatmap/?bookmark=92db485580786d1ef14c6c06b680201b