Published in final edited form as:
*Nat Biotechnol*. 2020 June ; 38(6): 675–678. doi:10.1038/s41587-020-0546-8.

# Visualizing and interpreting cancer genomics data via the Xena platform

**Mary J. Goldman**[1,8], **Brian Craft**[1,8], **Mim Hastie**[2], **Kristupas Repečka**[3], **Fran McDade**[2], **Akhil Kamath**[4], **Ayan Banerjee**[5], **Yunhai Luo**[6], **Dave Rogers**[2], **Angela N. Brooks**[7,1], **Jingchun Zhu**[1], **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]National Institute of Technology, Durgapur, India.

[6]Department of Genetics, Stanford University, Stanford, CA, USA.

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

## To the Editor

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)[1] and the GDC (Genomic Data Commons)[2], as well as smaller-scale datasets generated by individual labs. Commonly used interactive visualization tools are either web-based portals or desktop applications. Data portals have a dedicated back end and are a powerful means of viewing centrally hosted resource datasets (for example, Xena's predecessor, the University of California, Santa Cruz (UCSC) Cancer Browser (currently retired[3]), cBioPortal[4], ICGC (International Cancer Genomics Consortium) Data Portal[5], GDC Data Portal[2]). However, researchers wishing to use a data portal to explore their own data have to either redeploy the entire platform, a difficult task even for bioinformaticians, or upload private data to a server outside the user's control, a non-starter for protected patient data, such as germline variants (for example, MAGI (Mutation Annotation and Genome Interpretation[6]), WebMeV[7] or Ordino[8]). Desktop tools can view a user's own data securely (for example, Integrated Genomics Viewer (IGV)[9], Gitools[10]), but lack well-maintained, prebuilt files for the ever-evolving and expanding public data resources. This dichotomy between data portals and

mary@soe.ucsc.edu.

[8]These authors contributed equally: Mary Goldman, Brian Craft.

Competing interests
The authors declare no competing interests.

Additional information
**Supplementary information** is available for this paper at https://doi.org/10.1038/s41587-020-0546-8.

desktop tools highlights the challenge of using a single platform for both large public data and smaller-scale datasets generated by individual labs.

Complicating this dichotomy is the expanding amount, and complexity, of cancer genomics data resulting from numerous technological advances, including lower-cost high-throughput sequencing and single-cell-based technologies. Cancer genomics datasets are now being generated using new assays, such as whole-genome sequencing[11], DNA methylation whole-genome bisulfite sequencing[12] and ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing[13]). Visualizing and exploring these diverse data modalities is important but challenging, especially as many tools have traditionally specialized in only one or perhaps a few data types. And although these complex datasets generate insights individually, integration with other omics datasets is crucial to help researchers discover and validate findings.

UCSC Xena was developed as a high-performance visualization and analysis tool for both large public repositories and private datasets. It was built to scale with the current and future data growth and complexity. Xena's privacy-aware architecture enables cancer researchers of all computational backgrounds to explore large, diverse datasets. Researchers use the same system to securely explore their own data, together or separately from the public data, all the while keeping private data secure. The system easily supports many tens of thousands of samples and has been tested with up to a million cells. The simple and flexible architecture supports a variety of common and uncommon data types. Xena's Visual Spreadsheet visualization integrates gene-centric and genomic-coordinate-centric views across multiple data modalities, providing a deep, comprehensive view of genomic events within a cohort of tumors.

UCSC Xena (http://xena.ucsc.edu) has two components: the front end Xena Browser and the back end Xena Hubs (Fig. 1). The web-based Xena Browser empowers biologists to explore data across multiple Xena Hubs with 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 can be public or private (Supplementary Fig. 1). The Xena Browser receives data simultaneously from multiple Xena Hubs and integrates them into a single coherent visualization within the browser.

A private Xena Hub is a hub installed on a user's own computer (Supplementary Fig. 2). It is configured to only respond to requests from the computer's localhost network interface (that is, http://127.0.0.1). This ensures that the hub only communicates with the computer on which the hub is installed. A public hub is configured to respond to requests from external computers. There are two types of public Xena Hubs (Supplementary Fig. 2). The first type is an open-public hub, which is a public hub accessible by everyone. While we host several open-public hubs (Supplementary Table 1), users can also set up their own as a way to share data. An example of one is the Treehouse Hub set up by the Childhood Cancer Initiative to share pediatric cancer RNA-seq gene expression data (Supplementary Note). The second type of public hub is firewall-protected-public hub, which is a system in which, while the hub is configured to respond to requests from external computers, access to the computer itself is controlled via a firewall or similar technology. This model takes advantage of security that is typically already in place, thereby reducing user workload by not requiring reauthorization. Users who host this type hub will share the URL with authorized individuals. Other users who may inadvertently acquire knowledge of the protected hub will not be able to connect because of the firewall. This second type of public Xena Hub is useful for sharing private data within a lab or institution, as everyone who already has access to the data via the firewall will have access to the Xena Hub.

Public and private Xena Hubs use the same software; the only difference is in their configuration. Hubs default to only respond to requests from the computer's localhost network, locking down data accessibility to the host computer. Hubs only respond to external network requests if a user configures the hub to do so. Xena Hubs are designed to be turnkey, allowing users who may not be computationally savvy to easily install and use a Xena Hub on their personal computer (https://xena.ucsc.edu/private-hubs/). An interactive setup wizard guides users through the process of installing and running a Xena Hub on their Windows or Macintosh computers, while a web wizard guides them through the data loading process (https://bit.ly/localUCSCXenaHub). In addition to these user-friendly wizards, Xena Hubs can be installed and used via the command line on Windows, Macintosh and Linux machines.

The Xena Browser automatically connects to a default list of the open-public hubs we host (Supplementary Table 1) and, if it exists, to the private local hub on users' computer. Users can add a new hub by entering the hub URL on the Xena Browser Data Hubs page. Data integration occurs only within the Xena Browser, keeping private data secure. Genomic data flows from a Xena Hub to the Xena Browser (Fig. 1), which never communicates the genomic or phenotypic data it displays back to any hub or server. The only exception to this model occurs when saving bookmarks as URLs, a feature that allows users to save live views of their current visualization. If a visualization contains only data from the list of open-public Xena Hubs we host, users can generate a URL for their current view, which will take researchers back to the live browser session. As the data are already public, we store the data in the view for each URL on our web server, allowing them to be shared with colleagues or included in presentations. If a view contains any data from any other type of hub, users are instead required to download the current visualization 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 shared (for example, via e-mail) and then imported back into the Xena Browser to recreate the live browser session. Thus, even when the user creates bookmarks, protected data are kept secure through Xena's architecture and the use of private hubs.

Xena's architecture of a decoupled Xena Browser and multiple Xena Hubs enables several other features. First, researchers can easily view their own private data by installing their own Xena Hub on their computer. Xena Hubs are lightweight compared with a full-fledged application and install easily. Second, users can use the same platform to view public and private data together. Xena integrates data across multiple hubs, allowing users to view data from separate hubs as a coherent data resource (Fig. 2). Xena does this while keeping private data on the user's own computer and not uploading it to a public server while also avoiding the need to download large public resources. This is especially useful for researchers who wish to view their own analysis results on public data, such as their own clustering calls, but do not want to host a separate version of these 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. These advantages, in addition to the security advantages, are a major departure from and innovation over the UCSC Cancer Browser.

The Xena Browser (https://xenabrowser.net) has a wide variety of visualizations and analyses, including survival analyses, scatter plots, bar graphs, statistical tests and genomic signatures, as well as our Visual Spreadsheet view. The Xena Visual Spreadsheet was designed to enable and enhance integration across diverse data modalities, providing researchers a more biologically complete understanding of genomic events and tumor biology. Analogous to an office spreadsheet application, it is a visual representation of a data grid in which each column is a slice of genomic or phenotypic data (for example, gene expression, mutation calls, methylation probes, subtype classifications or age) and each row is a single entity (for example, a bulk tumor sample, cell line or single cell; Fig. 2). Xena's Visual Spreadsheet displays genomic data in a wide variety of gene-centric, coordinate-centric and feature-centric views (Supplementary Fig. 3) for both coding and non-coding regions (Supplementary Fig. 4). Dynamic web links to the UCSC Genome Browser give genomic context to any gene, chromosome region or feature. Researchers can easily reorder the Visual Spreadsheet, hierarchically cluster genes, and zoom in to just a few samples or out to the whole cohort, all leading to a near-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 (Fig. 2).

The power of the Visual Spreadsheet is its deep data integration. 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 whether higher expression for a gene is driven by copy-number amplification or by a missense mutation (Supplementary Fig. 5) or by demethylation and opening of the promoter region, as reflected in DNA methylation and ATAC-seq data (Supplementary Fig. 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 whether a gene amplification is part of a chromosomal arm duplication or a focal amplification (Supplementary Fig. 6). 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 whether increased homologous recombination deficiency (HRD) signature scores are enriched in a specific cancer type or subtypes (Supplementary Fig. 7). Finally, integration across a user's own data and public resources on the same samples helps users to gain insights into their own data. 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 (Fig. 2). Because Visual Spreadsheet does not differentiate between public data and private data from a rendering perspective, it appears to the user that all data come from a coherent source. These diverse integrations help researchers harness the power of comprehensive genomics studies — either their own or from public resources — driving discovery and a deeper understanding of cancer biology.

In addition to the Visual Spreadsheet, Xena has many other powerful views, analyses and functionalities. Our powerful text-based search allows users to dynamically highlight, filter and group samples (Supplementary Fig. 8). Researchers use this to search the data on the screen similarly to using the 'find' functionality in Microsoft Word. Samples are matched and highlighted in real time as the user types. Researchers can then filter to their samples of interest or dynamically build subgroups. This is a powerful way to dynamically construct subpopulations based on any genomic data for comparison and analysis. Xena also has highly configurable Kaplan–Meier analyses, bar charts, box plots and scatter plots, all with statistical tests automatically computed (Supplementary Figs. 7 and 9). We support data sharing through bookmarked views and high-resolution PDFs. Genomic signatures are easily built over gene expression data or any other genomic data type.

Performance is critical for interactive visualization tools, especially on the web. Growing sample sizes for genomic experiments have become a challenge for many tools, including for the UCSC Cancer Browser. Knowing this, we optimized Xena to support visualizations on many tens of thousands of samples, delivering slices of data in milliseconds to a few seconds. During the first eight months in 2019, we averaged 1,400 users a week with an average concurrent current usage of 3.34 users. To ensure we will continue to perform as we scale, we tested our public hubs deployed in the cloud with 50 concurrent requests and had an average response rate of 244 ms.

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 ATAC-seq peak signals (Supplementary Fig. 3) and structural variant data (Supplementary Fig. 10) — a substantial advantage over the UCSC Cancer Browser. Current supported data modalities include somatic and germline single nucleotide polymorphisms (SNPs); indels; large structural variants; copy-number variation; gene, transcript, exon, protein or miRNA expression; DNA methylation; ATAC-seq peak signals; phenotypes; clinical data; and sample annotations[14,15].

UCSC Xena provides interactive online visualization of seminal cancer genomics datasets through multiple open-public Xena Hubs. We host over 1,600 datasets from over 50 cancer types, including the latest from TCGA, ICGC, TCGA Pan-Cancer Atlas[16], PCAWG (Pan-Cancer Analysis of Whole Genomes)[11] and the GDC (Supplementary Table 1). Xena Hubs offer a performance advantage over these resources' native APIs (application programming interfaces), especially when visualizing more than just a few samples. We use custom ETL (extract–transform–load) processes to keep the Xena Hubs updated with the latest data from their respective sources (Supplementary Fig. 1). We only download and process the derived datasets from each source, such as gene expression values, leaving the raw sequencing data at their respective locations. Xena complements each of these resources by providing powerful interactive visualizations for these data.

In addition to these well-known resources, we also host results from the UCSC Toil RNA-seq recompute compendium, a uniformly realigned and re-called gene and transcript expression dataset for all TCGA, TARGET (Therapeutically Applicable Research to Generate Effective Treatments) and GTEx (Genotype–Tissue Expression) samples[17]. This dataset allows users to compare gene and transcript expression of TCGA tumor samples to corresponding GTEx normal samples[18]. The UCSC public hub hosts data curated from various publications.

UCSC Xena complements existing tools, including the cBioPortal, ICGC Portal, GDC Portal, IGV and St. Jude Cloud[19], in several ways. First, our focus is 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 differentiates itself from other tools by enabling researchers to view their own data together with consortium data while still maintaining privacy. Furthermore, Xena focuses on integrative visualization of multi-omics datasets across different genomic contexts, including genes, genomic elements or any genomic region, for both coding and non-coding parts 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 from 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 this challenge.

Although it is widely recognized that data sharing is key to advancing cancer research, how it is shared can affect the ease of data access. UCSC Xena is 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 public Xena Hub on a server, in the cloud, or as part of an analysis pipeline, making generated data available in a user-friendly manner that requires little extra effort. Access control for public hubs shared with a limited set of researchers is currently only provided by protecting the computer itself, such as using a firewall. To fully support users who need to share private data but for whom firewall protection is inadequate or not feasible, we plan to develop hub-wide user authorization capability to allow the hub owner to precisely control who has access to the hub data. This would be useful for collaborative projects, which share data with users across multiple institutions. It will also allow integration with existing federated authentication and authorization services. Data sharing has advanced, 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 its web-browser-based visualization and separate data hubs, allows new projects to easily add their data to the growing public compendium. We support many different data modalities, both those existing now and those to come in the future, by maintaining flexible input formats. Xena excels at showing trends across cohorts of samples, cells or cell lines. Although 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 data sharing, integration and visualization needs of the cancer research community.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

Research reported in this publication was supported by the National Cancer Institute of the US 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 the Silicon Valley Community Foundation. We would also like to thank AWS Cloud Credits for Research and Google Summer of Code.

## References

1. Chin L, Hahn WC, Getz G & Meyerson M Genes Dev. 25, 534–555 (2011). [PubMed: 21406553]
2. Grossman RL et al. N. Engl. J. Med 375, 1109–1112 (2016). [PubMed: 27653561]
3. Zhu J et al. Nat. Methods 6, 239–240 (2009). [PubMed: 19333237]
4. Cerami E et al. Cancer Discov. 2, 401–404 (2012). [PubMed: 22588877]
5. Zhang J et al. Nat. Biotechnol 37, 367–369 (2019). [PubMed: 30877282]
6. Leiserson MDM et al. Nat. Methods 12, 483–484 (2015). [PubMed: 26020500]
7. Wang YE, Kutnetsov L, Partensky A, Farid J & Quackenbush J Cancer Res. 77, e11–e14 (2017). [PubMed: 29092929]
8. Streit M et al. Bioinformatics 35, 3140–3142 (2019). [PubMed: 30657871]
9. Thorvaldsdóttir H, Robinson JT & Mesirov JP Brief. Bioinform 14, 178–192 (2013). [PubMed: 22517427]
10. Perez-Llamas C & Lopez-Bigas N PLoS One 6, e19541 (2011). [PubMed: 21602921]
11. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Nature 578, 82–93 (2020). [PubMed: 32025007]
12. Zhou W et al. Nat. Genet 50, 591–602 (2018). [PubMed: 29610480]
13. Corces MR et al. Science 362, 6413 (2018).
14. Cieślik M & Chinnaiyan AM Nat. Rev. Genet 19, 93–109 (2018). [PubMed: 29279605]
15. Langmead B & Nellore A Nat. Rev. Genet 19, 208–219 (2018). [PubMed: 29379135]
16. Hoadley KA et al. Cell 173(291–304), e6 (2018).
17. Consortium GTEx et al. Nature 550, 204–213 (2017). [PubMed: 29022597]
18. Vivian J et al. Nat. Biotechnol 35, 314–316 (2017). [PubMed: 28398314]
19. Ma X et al. Nature 555, 371–376 (2018). [PubMed: 29489755]
20. Gao Q et al. Cell Rep. 23(227–238), e3 (2018).

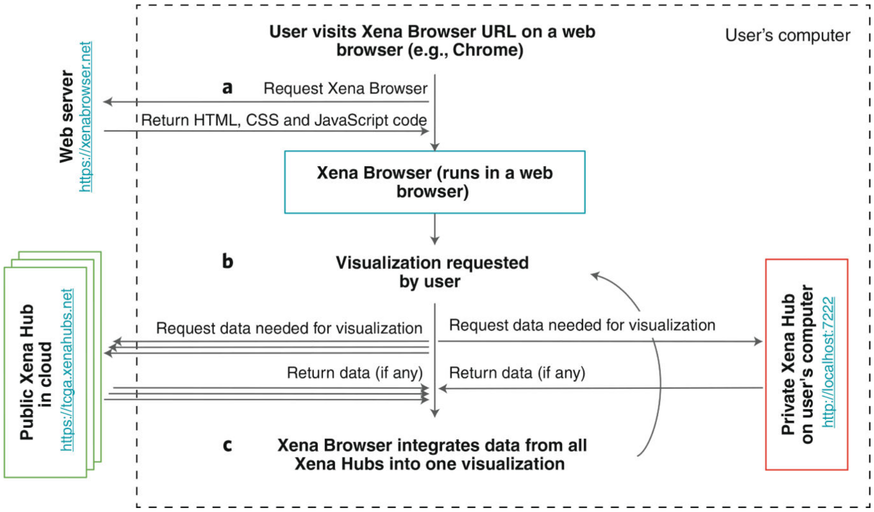
User visits Xena Browser URL on a web browser (e.g., Chrome)
User's computer
Web server
https://xenabrowser.net
a
Request Xena Browser
Return HTML, CSS and JavaScript code
Xena Browser (runs in a web browser)
b
Visualization requested by user
Public Xena Hub in cloud
https://tcga.xenahubs.net
Request data needed for visualization
Request data needed for visualization
Return data (if any)
Return data (if any)
Private Xena Hub on user's computer
http://localhost:7222
c
Xena Browser integrates data from all Xena Hubs into one visualization

**Fig. 1 |. Xena's architecture to securely join public and private data.**

Data always flow from the Xena Hubs to the Xena Browser for visualization and integration. **a**, The user's web browser (for example, Google Chrome) requests the Xena Browser code and runs it. **b**, using the Xena Browser, the user requests a visualization, initiating a request for data from the Xena Browser's list of public hubs. Simultaneously with this request, the Xena Browser requests data from the private local hub on the user's computer. **c**, The Xena Browser code combines data from all Xena Hubs together into one coherent visualization. The user can then interact with the visualization to trigger a new data request.

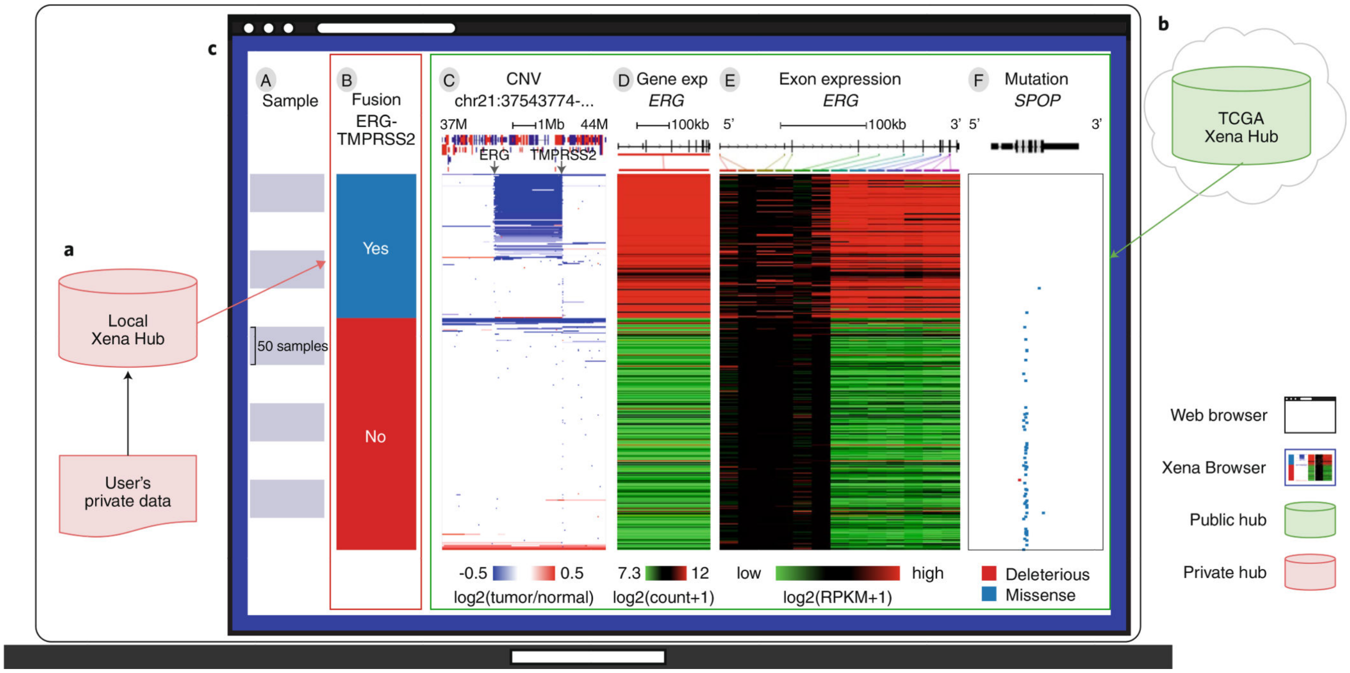

**Fig. 2 |. An example Xena Browser Visual Spreadsheet examining published ERG–TMPRSS2 (ETS transcription factor–transmembrane serine protease 2) fusion calls in TCGA PRAD (prostate cancer) by combining data from local and public Xena hubs.**

**a**, A user downloads *ERG-TMPRSS2* fusion calls on TCGA PRAD samples from Gao et al.[20] ($n = 492$) and loads the data into their own local Xena Hub. **b**, TCGA copy number, gene expression and mutation data from the same samples are available via the public TCGA hub. **c**, The user then compares the fusion calls to the public data using Xena Browser Visual Spreadsheet. Column B is the fusion call from Gao et al. Column C is copy number variation data, zoomed in to a region of chromosome 21 (37–44 Mb). Amplifications are in red and deletions are in blue. The diagram at the top shows genes along the chromosome; red genes are on the positive strand and blue are on the negative strand. Columns D is *ERG* gene expression and Column E is *ERG* exon expression. expression is colored red to green for high to low expression. The gene diagram at the top shows exons as boxes, with tall coding regions and shorter untranslated regions. Column F is *SPOP* (speckle type BTB/POZ protein) 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: mutations computationally predicted to truncate a transcript, such as frameshift or nonsense mutations, are in red and missense mutations are in blue. We can see that the fusion calls are highly consistent with the characteristic overexpression of *ERG* (columns D and E). 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 intrachromosomal deletion of chromosome 21 that fuses *TMPRSS2* to *ERG*, as shown in column C. This observation is consistent with the 63.3% validation rate described in Gao et al.[20]. *SPOP* mutations (blue tick marks in column F) are mutually exclusive with the fusion event. Rows are sorted by the leftmost data column (column B) and sub-sorted on columns thereafter.