Online resources for PCAWG data exploration, visualization, and discovery

Mary Goldman*[1], Junjun Zhang*[2], Nuno A. Fonseca*[3], Qian Xiang[2], Brian Craft[1], Elena Piñeiro-Yáñez[4], Brian D O'Connor[1], Wojciech Bazant[3], Elisabet Barrera[3], Alfonso Muñoz[3], Robert Petryszak[3], Anja Füllgrabe[3], Fatima Al-Shahrour[4], Maria Keays[3], David Haussler[1], John N. Weinstein[5], Wolfgang Huber[6], Alfonso Valencia[7], Irene Papatheodorou#[3], Jingchun Zhu#[1], Vincent Ferreti#[2], Miguel Vazquez#[8], on behalf of the PCAWG Portals and Visualization Working Group, and the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network

[1]UC Santa Cruz Genomics Institute, Santa Cruz, 95064, USA; [2]Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada; [3]European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Hinxton, CB10 1SD, UK; [4]Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), Madrid , 28029, Spain; [5]UT MD Anderson Cancer Center, Houston, 77030, USA; [6]European Molecular Biology Laboratory, Heidelberg, 69117, Germany; [7]Barcelona Supercomputing Center (BSC), Barcelona, 08034, Spain **&** ICREA, Barcelona, 08010, Spain; [8]Barcelona Supercomputing Center (BSC), Barcelona, 08034, Spain & Norwegian University of Science and Technology, Trondheim, Norway.

Abstract

The Pan-Cancer Analysis of Whole Genomes (PCAWG) cohort provides a large, uniformly-analyzed, whole-genome dataset. The PCAWG Landing Page (http://docs.icgc.org/pcawg) focuses on four biologist-friendly, publicly-available web tools for exploring this data: The ICGC Data Portal, UCSC Xena, Expression Atlas, and PCAWG-Scout. They enable researchers to dynamically query the complex genomics data, explore tumors' molecular landscapes, and include external information to facilitate interpretation.

Paper

The Pan-Cancer Analysis of Whole Genomes (PCAWG) cohort provides the largest uniformly-analyzed, publicly-available whole-genome dataset in cancer genomics. The PCAWG project has reprocessed whole-genome shotgun-sequencing data for 2,658 donors from 48 cancer projects across 20 different primary sites and 14 jurisdictions. This reprocessing involved alignment to the reference genome, quality assessment, calling somatic and germline variants, and generation of consensus mutation calls derived from multiple bioinformatics methods (Campbell 2017, Yung 2017). From the PCAWG Landing Page (http://docs.icgc.org/pcawg) (Supplementary Fig. 1), users can find four biologist-friendly, publicly-available web resources developed for downloading, visualizing, and exploring this data: The ICGC Data Portal, UCSC Xena, Expression Atlas, and PCAWG-Scout (Supplementary Table 1). These resources enable biologists to dynamically query and integrate the complex genomics data, explore tumors' molecular landscapes, and combine external data with the PCAWG cohort to facilitate interpretation (Table 1).

These resources support the primary 'omic data types generated by the PCAWG project, including simple somatic mutations (single- and multiple- nucleotide variants (SNVs, MNVs)); small insertions and deletions (INDELs); large somatic structural variants (SVs); copy number variants; gene fusions; RNAseq gene- and miRNA-expression; DNA methylation; and phenotypic annotations (Li 2017,

PCAWG Transcriptome Core Group 2018, Rheinbay 2017, Sabarinathan 2017, Yung 2017). Two types of files were generated by the PCAWG analysis: primary BAM and VCF files, and downstream analysis results (Supplementary table 2) (Campbell 2017). The ICGC Data Portal provides a uniform search interface for both all file types (Yung 2017). Each of the three resources, UCSC Xena, Expression Atlas, and PCAWG-Scout, separately ingested the same primary result files and individually refined them for online visualization, exploration and download (Fig. 1A).

In addition to the ICGC Data Portal serving as the main entry point for accessing PCAWG data, users can also explore PCAWG consensus simple somatic mutations by their frequencies, patterns of co-occurrence, mutual exclusivity, and functional associations. UCSC Xena integrates diverse types of genomic and phenotypic/clinical information at the sample-level across large number of samples, enabling visualizing patterns in the genomics data. Expression Atlas focuses on RNAseq data, supporting queries in either a baseline context (e.g., find genes that are expressed in PCAWG prostate adenocarcinoma samples) or in a differential context (e.g., find genes that are under- or over-expressed in prostate adenocarcinomas compared to “adjacent-normal” prostate samples). PCAWG-Scout allows users to run their own analysis on-demand. The wide variety of PCAWG-Scout methods include predicted cancer driver genes, differential gene expression, recurrent structural variation, survival analysis, pathway enrichment, protein structure, mutational signatures, and predictions of recommended therapies based on their in-house resource, PanDrugs (Supplementary Fig. 5).

Access to protected data (i.e., primary BAM and VCF files, germline variant calls, and simple somatic mutation calls for non-coding regions) is provided through the ICGC Data Portal for researchers with approved authorization (please see http://docs.icgc.org/portal/access/ on how to obtain access). UCSC Xena provides secure views of protected, non-coding somatic mutations through the use of a private Xena data hub. Both the ICGC Data Portal and PCAWG-Scout publicly display non-identifiable analysis results of protected data.

	Functionality	**ICGC Data Portal**	**PCAWG -Scout**	**UCSC Xena**	**Expression Atlas**
Search	Search by demographic data, specimen phenotype, and molecular subtype	Y	Y	Y	
	Search for genes and/or variants	Y	Y	Y	Y
	Search by genomic coordinates	Y		Y	
Visualize	Visualize a cohort of samples	Y	Y	Y	Y
	Visualize multiple types of data on sample-level		Y	Y	
	Visualize coding mutations	Y	Y	Y	
	Visualize non-coding variants	Y		Y	
	Visualize structural variants		Y	Y	

	Visualize mutational signatures and predicted drivers		Y		
	Visualize tissue expression on a human figure				Y
	Visualize gene co-expression			Y	Y
	Visualize pathways and/or therapeutic associations	Y	Y		
	Visualize summary of BAMs / VCFs	Y			
Analysis	Discover differentially- or co-expressed genes, and/or mutually exclusive genomic events		Y		
	Perform geneset / pathway enrichment analysis	Y	Y		Y
	Perform Kaplan-Meier plot with survival statistics	Y	Y	Y	
	Access public analysis results of protected data	Y	Y		
Download	Download BAMs, VCFs, working group results	Y			
	Download secondary processed data	Y	Y	Y	Y
	Programmatic download of data	Y	Y	Y	Y
Integration	Integrate with non-PCAWG genomics data	Y	Y	Y	Y
	Integrate external annotations from COSMIC, the UCSC Genome Browser, Ensembl, pathway databases, drug target compendia, etc.	Y	Y	Y	

Table 1. Search, analysis, visualization, download and integration functionalities provided by the PCAWG online data resources.

ICGC Data Portal (https://dcc.icgc.org). The ICGC Data Portal serves as the main entry point for accessing PCAWG datasets with a single uniform web interface and a universal data download client. This uniform interface gives users easy access the myriad of PCAWG sequencing data and variant calls that reside in many repositories and compute clouds worldwide. The intuitive faceted search interface is enabled through permanent, unique ICGC identifiers for each file and a set of harmonized metadata (e.g. data types and formats, experimental assays and computation methods). iobio web socket streaming technology (Chase 2014) gives users high-level visualizations in real time of BAM and VCF files stored remotely on AWS and Collab. PCAWG consensus simple somatic mutations (excluding US projects due policy constraints) are integrated with clinical data elements and rich functional annotations including affected proteins, pathways, gene ontology terms, and other factors. The Advanced Search and Analysis tools allow users to explore functional associations with phenotypic data such as molecular subtype and patient survival. All other downstream analysis results generated by PCAWG working groups are available for download through the portal.

UCSC Xena (https://pcawg.xenahubs.net). UCSC Xena visualizes all PCAWG primary results, including copy number, gene expression, gene fusion, promoter usage, simple somatic mutations, large somatic structural variation, mutational signatures and phenotypic data. This open-access data is available through a public Xena hub ('PCAWG hub', Supplementary Table 3), while consensus simple somatic mutations (including both coding and non-coding) can be loaded into a user's local computer private Xena hub (Supplementary Fig. 2). The UCSC Xena Browser accesses data from multiple data hubs simultaneously, allowing users to visualize PCAWG data alongside their own private data while still maintaining data security. The Xena Browser excels at integrating simple mutations, structural variants, gene expression data, and more, for the same or multiple genes across many samples for large cohorts (Fig. 1B). Xena displays somatic mutations on genomic coordinates, transcripts, and 3D protein structures. Histograms, boxplots and scatterplots offer additional views and statistical analyses while Kaplan-Meier plots show survival analyses.

Expression Atlas (https://www.ebi.ac.uk/gxa/home). Expression Atlas contains in-house analysed RNAseq and expression microarray data for querying gene expression across tissues, cell types, developmental stages and/or experimental conditions (Petryszak 2016). Queries can be either in a baseline context or in a differential context. PCAWG RNAseq gene expression data (Fonseca 2017) are manually curated to a high standard by Expression Atlas curators (Supplementary Fig. 3) and are presented in a heatmap with summarized baseline expression. Two different views of the data are provided: summarized expression levels for each tumor type and gene expression at the level of individual samples. Both views include reference gene expression data sets for matching tissues from the GTEx study (Melé 2015) and any available adjacent normals from the PCAWG study (Figure 1C).

PCAWG-Scout (http://pcawgscout.bsc.es/)**.** PCAWG-Scout leverages the Ruby bioinformatics toolkit (Rbbt), a framework for 'omics workflow and website templating, to make on-demand, in-depth analyses over the PCAWG data openly available to the whole research community. Views of protected data are available that still safeguard sensitive data (Supplementary Fig. 4). Through the PCAWG-Scout web interface, users can access an array of reports and visualizations that leverage on-demand bioinformatic computing infrastructure to produce results in real-time, allowing users to discover trends as well as form and test hypotheses. The web interface and underlying infrastructure are open-source and can be installed locally, with new reports added or altered easily through the modular templating system. This also allows the entire analysis suite to be applied to datasets outside those from PCAWG.

Case study. We demonstrate the value of each resource using a common driver event in prostate cancer, fusion of the oncogene ERG (St. John 2012, Adamo 2016). Xena's visual spreadsheet shows that 8 out of the 18 PCAWG prostate donor samples with both whole-genome sequencing and RNAseq data harbor an ERG fusion. These donors also show ERG over-expression (Fig. 1B). A view of the PCAWG structural variant data shows that all fusion breakpoints are located at the start of ERG, leaving the ERG coding region intact while fusing to the promoter region of TMPRSS2 or SLC45A3 (Fig. 1B). Both TMPRSS2 and SLC45A3 are highly expressed in normal prostate tissues and prostate tumors, as shown in the Expression Atlas Baseline Expression Widget (Fig. 1C). Combined analysis of the PCAWG and GTEx datasets leads to the hypothesis that a subset of

prostate cancers, through genome rearrangement, hijack the promoters of androgen-responsive genes to increase ERG expression, resulting in an androgen-dependent over-expression of ERG.

Although the ERG fusions are frequent, 46% (89 out of 195) of the PCAWG prostate tumors do not show them (Supplementary Fig. 6). PCAWG-Scout's mutual exclusivity analysis shows that simple mutations in FOXA1, SPOP, SYNE1, and ANKFN1 are significantly associated with these non-fusion tumors (Fig. 1D). Furthermore, in PCAWG-Scout's 3D protein structure view, the mutations in SPOP are shown to cluster tightly around the interaction interface for PTEN (Fig. 1D), suggesting that those mutations may lead to altered protein function for SPOP.

Code availability and embeddable web modules. In addition to providing online resources for PCAWG data exploration, we have made the code for each resource publicly available (Supplementary Table 4). Further, embeddable javascript modules are available for ICGC's OncoGrid, Xena Visual Spreadsheet, and Expression Atlas Heatmap and Anatomogram Widget (Supplementary Table 5). These modules are open-source and can be incorporated into 3rd party web applications.

Other online PCAWG resources In addition to these four resources, there are additional online tools from the PCAWG consortium. Examples of this include the panorama of driver mutations in PCAWG tumors that can be explored via prepared Gitools interactive heatmaps (http://www.gitools.org/pcawg), browsed in IntOGen, at http://www.intogen.org/pcawg (Radhakrishnan 2017), and the Chromothripsis Explorer, which visualizes genomic rearrangements across PCAWG tumors (reference pending). We envision the PCAWG data will be incorporated into many other existing tools, and new resources will be developed.

Conclusion. The powerful, comprehensive, online resources described here facilitate exploration of the PCAWG datasets. They enable both bench biologists and bioinformaticians to gain insights from the data and make new discoveries.

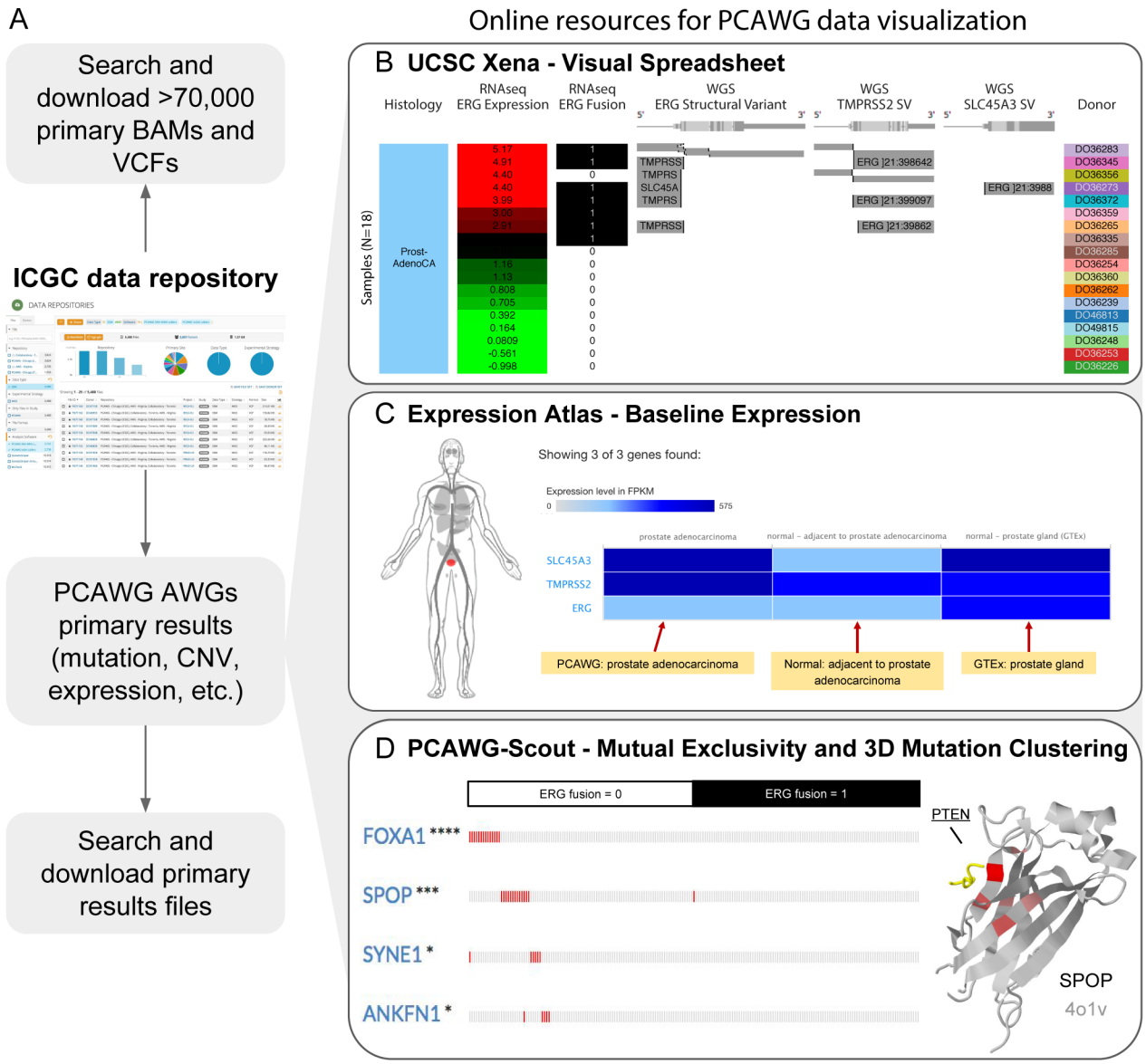


Figure 1, Online resources for PCAWG data download and visualization. These resources provide complementary views and analyses, illustrated by the example of mutually exclusive, recurrent events (such as ERG fusions and mutations in SPOP) detected in the prostate adenocarcinoma cohort.

A, All PCAWG BAMs, VCFs, and analysis working group primary results files are available through the *ICGC data portal*'s Data Repository tool (http://goo.gl/4ny2aG). To obtain these files, the user downloads a manifest of selected files and then downloads the actual data files (with authorization if needed) using the ICGC download tool. UCSC Xena, Expression Atlas and PCAWG-Scout have each pre-processed the same primary analysis working group results files. These tools's internal databases are designed for the fast response times needed for online applications.

B, The *Xena Visual Spreadsheet* shows that the ERG fusion is present in 8 out of 18 PCAWG prostate adenocarcinoma samples (https://goo.gl/KU333Z), as detected by the PCAWG RNAseq and whole-genome sequencing data. Each row corresponds to a sample. Columns, starting at the left, correspond to histology, ERG gene expression, and ERG fusion based on RNAseq data (merged fusion calls from two methods (PCAWG Transcriptome Core Group 2018). The next three columns show structural variant calls (consensus calls by four methods Li 2017) using whole genome DNAseq data for ERG, TMPRSS2, and SLC45A3. These

data show that TMPRSS2 and SLC45A3 are fusion partners for ERG, and that the fusions correlate with over-expression of ERG. Fusions detected by RNAseq and whole-genome sequencing are not always consistent. Here, even using a consensus of 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.

C, The *Expression Atlas* shows a heatmap of genes (rows) and tissue or disease type (columns). Here we show the expression of ERG, SLC45A3, and TMPRSS2 in healthy human tissue (top heatmap), as derived from our re-analysis of the GTEx data set. The bottom heatmap shows expression in PCAWG data (https://goo.gl/Ch1meK). The human figure, also know as an anatomogram, shows the highlighted prostate tissue. We can see that TMPRSS2 and SLC45A3 show very high prostate-specific expression. Both genes have been used by the tumor to fuse its promoter to ERG, driving ERG over-expression.

D, *PCAWG-Scout* complements the analysis by identifying recurrent mutational events in tumors without ERG fusion (fusion = 0). On the left, PCAWG-Scout shows a mutation exclusivity analysis (using Fisher's exact test), which identifies FOXA1 (****, < 0.0001), SPOP (***, < 0.001), SYNE1 (*, < 0.05), and ANKFN1 (*, < 0.05) as significantly associated with non-fusion tumors (https://goo.gl/nPJ45l). In the 3D protein structure of SPOP shown on the right, mutations are seen to cluster tightly around the region that overlaps with the interaction surface of PTEN. The portion of PTEN that interacts with SPOP is shown in yellow, along with the SPOP structure. Red indicates recurrent mutations in SPOP with a brighter red indicating higher rate of recurrence.

Methods to reproduce views in **A, B, C and D** are available in the Supplementary Information.

References

Adamo, P., Ladomery, M. R. The oncogene ERG: a key factor in prostate cancer. *Oncogene* **35,** 403-414 (2016).

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).

Chase, A. M., Yi, Q., DiSera, T., D'Astous, B. & Marth G. T. bam.iobio: a web-based, real-time, sequence alignment file inspector. *Nature Methods* **11,** 1189-1189 (2014).

Li, Y., Roberts, N., Weischenfeldt, J., Wala, J. A., Shapira, O., Schumacher ,S., et al. Patterns of structural variation in human cancer. Preprint at https://www.biorxiv.org/content/early/2017/08/27/181339 (2017).

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., et al. The human transcriptome across tissues and individuals. *Science* **348,** 660-665 (2015).

PCAWG Transcriptome Core Group, Calabrese, C., Davidson, N. R., Fonseca, N. A., He, Y., et al. Genomic basis for RNA alterations revealed by whole-genome analyses of 27 cancer types. Preprint at https://www.biorxiv.org/content/early/2018/03/12/183889 (2018).

Petryszak, R., Keays, M., Tang, Y. A., Fonseca, N. A., Barrera, E., et al. Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research* **44,** D746-D752 (2016).

Rheinbay, E., Nielsen, M. M., Abascal, F., Tiao, G., Hornshøj, H., et al. Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. Preprint at https://www.biorxiv.org/content/early/2017/12/23/237313 (2017).

Sabarinathan, R., Pich, O., Martincorena, I., Rubio-Perez, C., Juul, M., et al. The whole-genome panorama of cancer drivers. Preprint at https://www.biorxiv.org/content/early/2017/12/23/190330 (2017).

St John, J., Powell, K., Conley-Lacomb, M. K., Chinni, S. R. TMPRSS2-ERG Fusion Gene Expression in Prostate Tumor Cells and Its Clinical and Biological Significance in Prostate Cancer Progression. *Journal of Cancer Science and Therapy* **4,** 94-101 (2012).

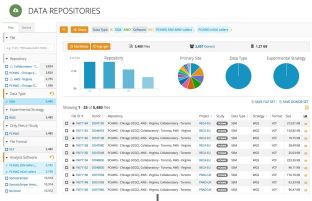
Yung C. K., O'Connor B. D., Yakneen, S., Zhang, J., Ellrott, K., et al. Large-Scale Uniform Analysis of Cancer Whole Genomes in Multiple Computing Environments. Preprint at https://www.biorxiv.org/content/early/2017/07/10/161638 (2017).

A

Online resources for PCAWG data visualization

Search and download >70,000 primary BAMs and VCFs

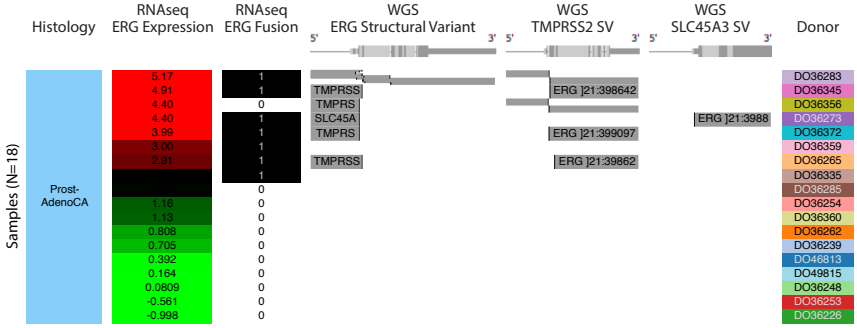
ICGC data repository



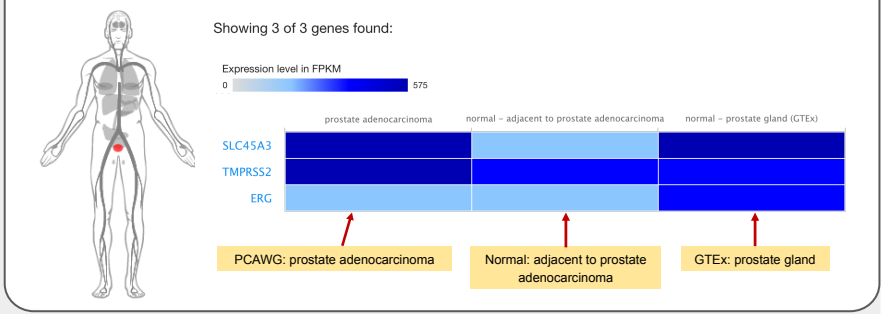
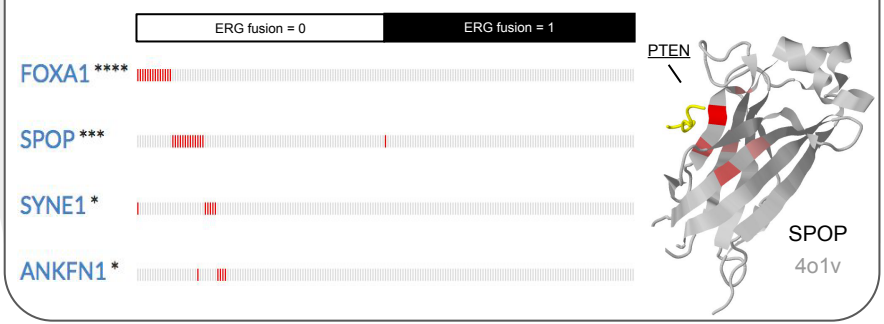
PCAWG AWGs primary results (mutation, CNV, expression, etc.)

Search and download primary results files

B UCSC Xena - Visual Spreadsheet



C Expression Atlas - Baseline Expression



D PCAWG-Scout - Mutual Exclusivity and 3D Mutation Clustering