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 Pineiro[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 Weinstein[5], Wolfgang Huber[6], Alfonso Valencia[7], Irene Papatheodorou#[3], Jingchun Zhu#[1], Vincent Ferreti#[2], Miguel Vazquez#[7], PCAWG-12 working group[8], PCAWG 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]Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), 28029 Madrid, Spain; [8] PCAWG-12 working group: Exploratory: portals, visualization and software infrastructure

Abstract

The PanCancer Analysis of Whole Genomes (PCAWG) cohort provides a large, uniformly-analyzed, whole-genome dataset. From the PCAWG Landing Page (http://docs.icgc.org/pcawg), users can access online resources for downloading, visualizing, and exploring this data: The ICGC Data Portal, UCSC Xena, Expression Atlas, and PCAWG-Scout. They enable researchers to dynamically query and integrate the complex genomics data, explore tumors' molecular landscapes, and combine external information with the PCAWG data to facilitate its interpretation.

Paper

The PanCancer 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 on 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 of all classes of somatic and germline variants, and generation of consensus mutation calls derived from multiple bioinformatics methods [cite marker paper, TECH paper]. From the PCAWG Landing Page (http://docs.icgc.org/pcawg) (Supplementary Fig. 1), users can access the four principal online resources developed for downloading, visualizing, and exploring this data: The ICGC Data Portal, UCSC Xena, Expression Atlas, and PCAWG-Scout (Supplementary Table 1). Together they serve the needs of bench biologists as well as expert bioinformaticians (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.

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), including copy number variants and gene fusions; RNAseq gene- and miRNA-expression; DNA methylation; and phenotypic annotations [cite various PCAWG papers]. Two types of files are generated by the PCAWG uniform analysis:

primary BAM and VCF files, and downstream analysis results (Supplementary table 2) [cite marker paper]. The ICGC Data Portal provides a uniform search interface for both types of files, which have been deposited in a variety of storage systems for long-term access [cite PCAWG tech group paper]. Each of the three resources, UCSC Xena, Expression Atlas, and PCAWG-Scout, separately ingest the same primary result files and individually refine them for online visualization, exploration and download (Fig. 1A).

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. UCSC Xena provides views of protected, non-coding somatic mutations that preserve data confidentiality through the use of a private data hub. Both the ICGC Data Portal and PCAWG-Scout publicly display non-identifiable analysis results for 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, pathway database, drug target compendia, etc.	Y	Y	Y	

Table 1. Search, analysis, visualization, download and integration functionalities provided by the PCAWG online data resources. *More information about therapeutic associations is available in Supplementary Figure 5.

ICGC Data Portal (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. Unique ICGC identifiers are assigned to each file for permanent referencing and a set of harmonized metadata (e.g. data types and formats, experimental assays and computation methods) are used to categorize the files, thereby enabling an intuitive faceted search interface. Real time BAM and VCF file random sampling enabled by iobio (Chase 2014) web socket streaming technology enables rapidly retrieval of high-level aggregation, and characteristics of sequence alignment and variant calling for data stored remotely on AWS and Collab. PCAWG consensus simple somatic mutations (excluding US projects due policy constraints) are loaded in the ICGC Data Portal where they are integrated with clinical data elements and rich functional annotations including affected proteins, pathways, gene ontology terms, and other factors. Via Advanced Search and Analysis tools users are able to explore frequencies of mutation events, patterns of co-occurrence and mutual exclusivity, as well as functional associations with phenotypic data such as molecular subtype and patient survival. Synapse pages with permanent, stable URLs are created to record metadata for all downstream analysis results generated by PCAWG working groups, with the actual data files being archived in ICGC Data Portal, available for download from the Data Release section.

UCSC Xena (https://pcawg.xenahubs.net). UCSC Xena provides data visualization for PCAWG copy number, gene expression, gene fusion, promoter usage, simple somatic coding mutations, large somatic structural variation, DNA methylation and phenotypic data through an open-access, publicly available Xena hub ('PCAWG hub'). Consensus simple somatic mutations (including both coding and non-coding mutations) from the PCAWG analysis [cite PCAWG-1 paper] can be loaded into a user's local computer private 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 privacy. The Xena Browser visualizes user-selected genomic data and associated phenotypic/clinical information using a spreadsheet view, similar to that of office

spreadsheet applications, that has been optimized for very large cohorts. Xena integrates multiple types of genomic data at the sample-level for an entire cohort. For example, it can simultaneously display simple mutations, structural variants and gene expression data for the same or multiple genes across many samples (Fig. 1B). Xena displays somatic mutations on genomic coordinates, transcripts, and 3D protein structures. Visualizations can be generated in a web browser for tens of thousands of samples. Histograms, boxplots and scatterplots offer additional views and statistical analyses. Kaplan-Meier plots show survival analyses. Detailed information on the data in the PCAWG hub (public) and user's local hub (private) is listed in Supplementary Table 3.

Expression Atlas (https://goo.gl/TsIYE5)**.** Expression Atlas contains analysed, in-house 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 (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 RNAseq gene expression data [cite PCAWG-3 paper] 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 (pcawgscout.bioinfo.cnio.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. Analyses include a wide array of methods such as predicted cancer driver genes, differential gene expression, recurrent structural variation, survival, pathway enrichment, protein structure, mutational signatures, and predictions of recommended therapies based on our in-house resource PanDrugs (Supplementary Fig. 5). 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 other than those from PCAWG.

Case study. We demonstrate the value of each resource using a common driver event in prostate cancer, ERG fusion (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. The 8 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 the ERG coding region, 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 drive ERG oncogene 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, the mutations in SPOP are shown through PCAWG-Scout's 3D protein view to cluster tightly around the interaction interface for PTEN (Fig. 1D), suggesting that those mutations may lead to altered protein function.

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

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). They are open-source, they are modular, and they can be incorporated into 3rd party web applications.

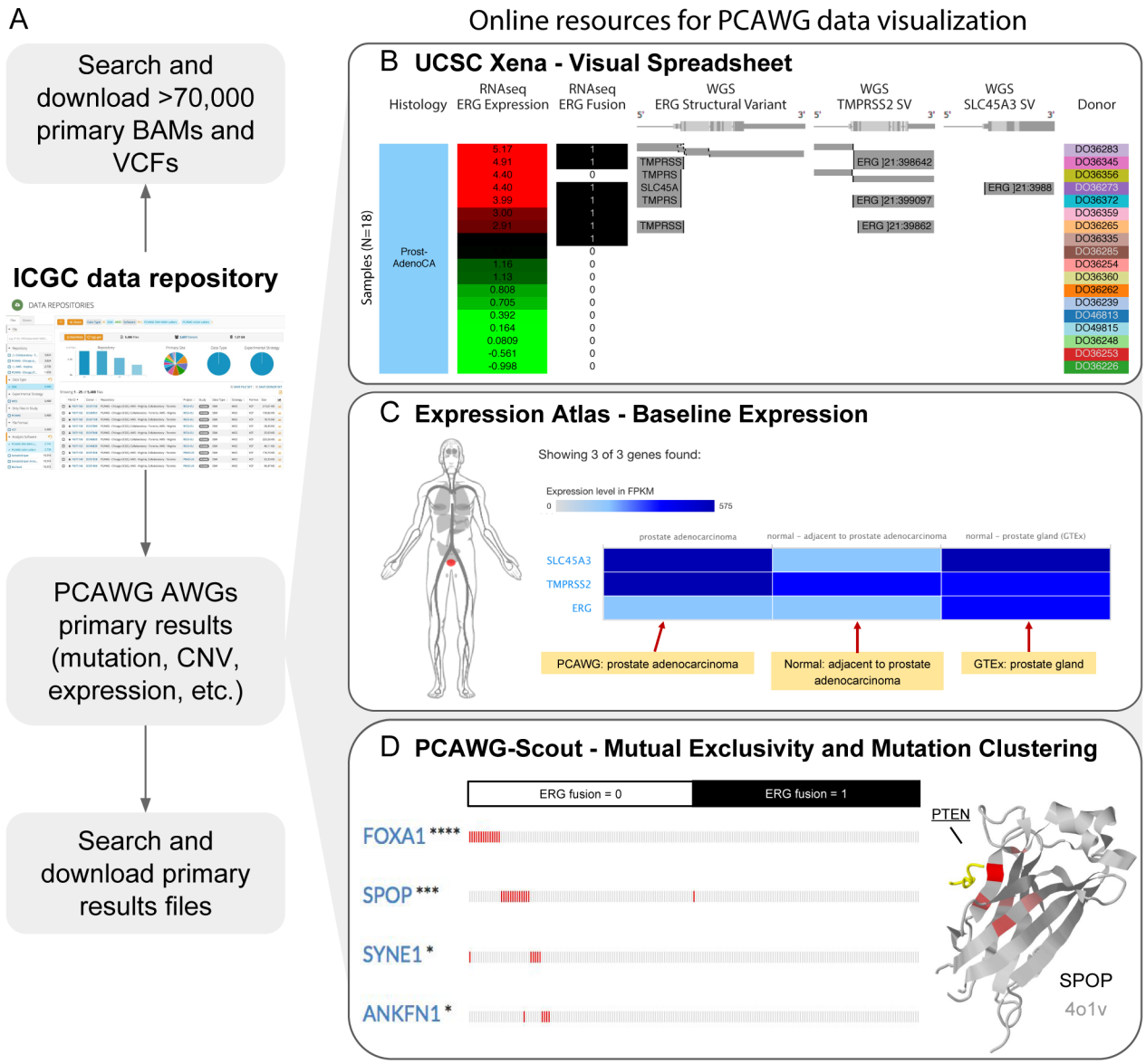

Figure 1 PCAWG data are available for download and visualization through multiple online resources. These resources provide complementary views of the data, illustrated by the example of mutually exclusive, recurrent events (such as ERG fusions and mutations in SPOP) detected in the prostate adenocarcinoma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

A, All PCAWG BAMs, VCFs and analysis working group primary results files are searchable through the *ICGC data portal*'s Data Repository tool (http://goo.gl/4ny2aG). To obtain these files, the user downloads a manifest of the 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 for storage in their internal databases, enabling the data to be visualized and explored quickly enough for online applications.

B, *Xena Visual Spreadsheet* shows that the ERG fusion detected in PCAWG RNAseq and whole-genome sequencing data is present in 8 out of 18 PCAWG prostate adenocarcinoma samples (https://goo.gl/auYmKX). 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 [cite PCAWG-3 gene fusion paper]). The next three columns show structural variant calls (consensus calls by four methods [cite PCAWG-6 paper]) using whole genome DNAseq data for ERG, TMPRSS2, and SLC45A3. The data show

that TMPRSS2 and SLC45A3 are fusion partners for ERG, and that 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 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.

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 (https://goo.gl/qe4vq7). The bottom heatmap shows expression in PCAWG data (pending url). The anatomogram (human figure) 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 to drive ERG over-expression.

D, *PCAWG-Scout* complements the analysis by identifying recurrent mutational events in tumors without ERG fusion (fusion = 0). To 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 structure shown on the right, SPOP 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 (yellow) along with the SPOP structure. Red indicates recurrent mutations in SPOP with 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 MR. (2016) The oncogene ERG: a key factor in prostate cancer. Oncogene. 35(4):403-14. doi: 10.1038/onc.2015.109. Epub 2015 Apr 27.

Chase AM, Yi Q, et al (2014) bam.iobio: a web-based, real-time, sequence alignment file inspector. *Nat Meth*, Vol. 11, No. 12, pp. 1189-1189.

Melé M, Ferreira PG, et al. (2015) The human transcriptome across tissues and individuals. Science. Vol 348 no. 6235 pp 660-665.

Petryszak R, Keays M, et al. (2016) Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants. Nucleic acids research Volume 44 p.D746-52

St John J, Powell K, Conley-Lacomb MK, Chinni SR. (2012) TMPRSS2-ERG Fusion Gene Expression in Prostate Tumor Cells and Its Clinical and Biological Significance in Prostate Cancer Progression. J Cancer Sci Ther. 4(4):94-101.

PCAWG marker preprint

PCAWG-TECH preprint

PCAWG-1 preprint: simple somatic mutations

PCAWG-2,5,9,14 preprint: driving mutations

PCAWG-3 preprint: gene expression

PCAWG-3 preprint: gene fusion

PCAWG-4 preprint: DNA methylation

PCAWG-6 preprint: structural variants

PCAWG-11 preprint: copy number

PCAWG-13/10 preprint: histology and subtype

PCAWG-14 preprint: miRNA