Supplemental Information of PCAWG-12 paper

Table of Contents

How to reproduce Figure 1

ICGC Data Repository (1A)

https://dcc.icgc.org/repositories?filters=%7B%22file%22:%7B%22study%22:%7B%22is%22:%5B%22PCAWG%22%5D%7D%7D%7D&files=%7B%22from%22:1%7D

Shortened url: http://goo.gl/4ny2aG

UCSC Xena Visual Spreadsheet (1B)

https://xenabrowser.net/heatmap/?bookmark=24ad428d0f3bf3bf3205bcffab64d276

Shortened url: https://goo.gl/auYmKX

Expression Atlas (1C)

The query in Figure 1C, showing PCAWG data for prostate adenocarcinoma together with adjacent normal tissue and normal prostate gland tissue from GTEx can be accessed through the link:
http://www.ebi.ac.uk/gxa/experiments/E-MTAB-5200/Results?specific=true&geneQuery=%255B%257B%2522value%2522%253A%2522TMPRSS2%2522%252C%2522category%2522%253A%2522symbol%2522%257D%252C%257B%2522value%2522%253A%2522ENSG00000157554%2522%257D%252C%257B%2522value%2522%253A%2522SLC45A3%2522%252C%2522category%2522%253A%2522symbol%2522%257D%255D&filterFactors=%257B%2522ORGANISM_PART%2522%253A%255B%2522prostate%2520gland%2522%255D%257D&cutoff=0.5&accessKey=1522478c-1bd0-4863-848e-d15e86774418

Shortened url: https://goo.gl/qe4vq7

PCAWG-Scout on demand analysis (1D)

Exclusivity analysis of non ERG fusion donors in PCAWG-Scout

To reproduce the exclusivity analysis we first need to produce the list of samples without ERG fusions which is done by generating the list of donors with ERG fusions and taking them out of all donors with SV (somatic structural variant) data.

To generate the list of ERG fusion donors in Prost-AdenoCa: (1) go to the report for Prost-AdenoCa (http://pcawgscout.bioinfo.cnio.es/entity/Study/Prost-AdenoCa); (2) click on 'SV summary' button (toward the bottom of the page); (3) in the resulting SV (structural variant) table, one can select fusions involving ERG by filtering the table, which is done by clicking the 'filter' button at the bottom of the table and then, in the popup window, typing 'ERG' in the field for 'Gene 1'; (4) after the table is filtered you can find the associated donors by selecting the 'Fusion donors' column, which is done by clicking the 'column' button at the bottom of the table

and then, in the popup window, clicking ‘save list’ button next to the label 'Fusion donors'. The list will open in a popup window. Open this report on the main window using the link button on the popup window header bar. You may rename the list using ‘Edit’ button on the sidebar of the report. For your convenience you may access the list **Prost-AdenoCa ERG fusion donors** (http://pcawgscout.bioinfo.cnio.es/entity_list/Sample/Prost-AdenoCa%20ERG%20fusion%20donors). Make it a favourite by clicking in the star icon on the top so we can use it later analyses.

We now need to complement this list: (1) go back to the report for Prost-AdenoCa (http://pcawgscout.bioinfo.cnio.es/entity/Study/Prost-AdenoCa) and click on the link for 'SV donors'; (2) from the sidebar, select the button ‘Compare'; (3) Since you made the previous list a favourite, you will see it now in the popup window, where you click the button ‘Remove’ to generate the complement donor list.

As before you may rename that list and save it as a favourite. For your convenience this list is accessible as **Prost-AdenoCa ERG non-fusion donors** (http://pcawgscout.bioinfo.cnio.es/entity_list/Sample/Prost-AdenoCa%20ERG%20non-fusion%20donors). You can now click on the ‘Characteristic alterations’ button (at the bottom of the page) to perform the exclusivity analysis. This will start the on-demand analysis, which will complete in less than a minute. The result is the list of gene alterations that are enriched in the non-fusion donors with associated statistical significance, as shown in **Figure 1D** (http://pcawgscout.bioinfo.cnio.es/entity_list_action/Sample/characteristic_alterations/Prost-AdenoCa%20ERG%20non-fusion%20donors)

Annotation and 3D clustering of SPOP mutations in PCAWG-Scout

To reproduce the image with the SPOP/PTEN structure, search for SPOP in the search box at the top of the page, and select **the first protein isoform** (http://pcawgscout.bioinfo.cnio.es/entity/Protein:Ensembl%20Protein%20ID/ENSP00000240327?organism=Hsa/feb2014) from the report sidebar. To show the protein report in the JMOL viewer, click the "JMOL" tab and select the PDB 4o1v. To display the mutation density gradient overlaid, click the "PCAWG" tab, then the ‘Highlight’ button, and go back to the "JMOL" tab to show this gradient with respect to all PCAWG donors.

To focus on prostate samples, first click the "PCAWG" tab, then click on ‘filter’, type "Prost-AdenoCa" in the field "histology_abbreviation", and click submit to filter to only those samples. Click ‘Highlight’ and then click on the "JMOL" tab to view SPOP mutations only from prostate samples.

You can use the ‘Sequence’ tab to visualize the SPOP mutation clustering on a linear depiction of protein sequences (not included in figure). For further confirmation click the ‘Protein feature incidence’ to see a binomial distribution analysis that detects if this region is both significantly mutated and annotated as ‘Important for binding substrate proteins’. Mutation F -> A in residue 133 (the most recurrently mutated in this cohort) is annotated as ‘Strongly reduced affinity for substrate protein’.

Supplementary information of PCAWG data wrangling carried out in Xena

PCAWG analysis working group primary results files (listed in Supplementary Table 2) are downloaded and wrangled into two Xena cohorts: (i) the "PCAWG donor-centric" cohort, where all datasets use ICGC Donor IDs, such as DO217962 and (ii) the "PCAWG specimen-centric" cohort, where all datasets use ICGC Specimen IDs, such as SP117136.

Several steps were taken to wrangle all genomics and phenotypic datasets into Xena. For genomics datasets, we mapped data from aliquot IDs to the donor IDs for the donor-centric cohort and to the specimen IDs for the specimen-centric cohort. Specific to the donor-centric cohort, data from normal specimens were removed, so that the data represents only profiles from the tumor. When multiple specimens are available for the same donor, an average was taken and assigned to the donor.

For phenotype data, we mapped the specimen histology classifications back to the donors IDs after data from normal specimens was removed. We propagated donor clinical data to all specimen IDs belonging to the donor.

We extracted coding mutations from the consensus simple somatic mutation datasets and made this derived dataset available on the open-access PCAWG hub (Consensus SNVs and indels - coding). The protected whole-genome consensus simple somatic mutation dataset are downloaded, wrangled into xena-ready format. The resulting xena ready file was uploaded back to PCAWG controlled data access storage (syn7122445), where the data remains under controlled access.

Supplementary Figure 1. PCAWG online resources landing page.

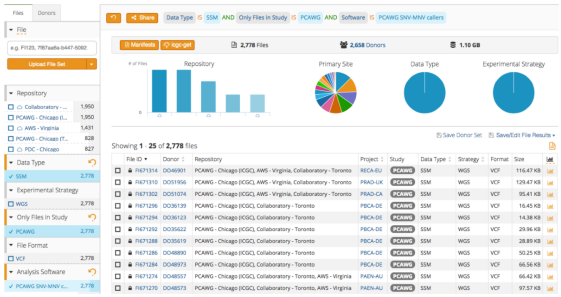
ICGC data portal provides a landing page for PCAWG data, highlighting the four online resources. From here, users can link to each individual resource's PCAWG page, which are listed in Supplemental Table 1.

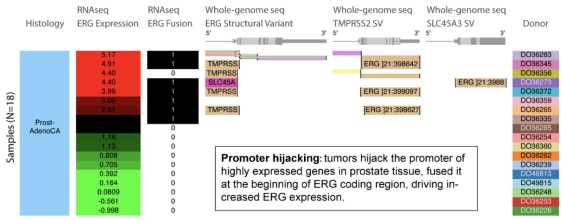
PCAWG Data Portal and Visualizations

The PCAWG study is an international collaboration to identify common patterns of mutation in more than 2,800 cancer whole genomes from the International Cancer Genome Consortium. The project produced large amount data with many types including simple somatic mutations (SNVs, MNVs and small INDELs), large-scale somatic structural variations, copy number alterations, germline variations, RNA expression profiles, gene fusions, and phenotypic annotations etc. PCAWG data have been imported, processed and made available in the following four major online resources for download and exploration by the cancer researchers worldwide.

ICGC Data Portal

The ICGC Data Portal is the main data dissemination platform for ICGC. PCAWG data have been imported into or indexed by the data portal which makes data search / download / exploration simple and effective. Explore ICGC Data Portal: PCAWG Somatic Mutations, PCAWG data files in various repositories.



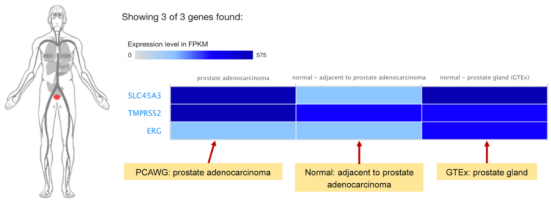


UCSC Xena

UCSC Xena is an exploration tool for multi-omic resource data, enabling discovery of correlations among all open-access PCAWG primary results, and performance of survival analyses. Explore PCAWG data in UCSC Xena at https://pcawg.xenahubs.net.

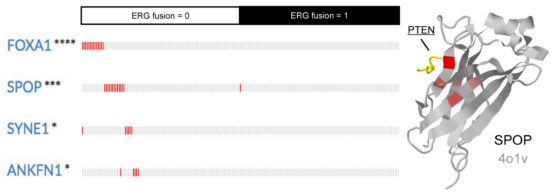
EBI Expression Atlas

Expression Atlas is an open science resource that gives users a powerful way to find information about gene and protein expression across species and biological conditions such as different tissues, cell types, developmental stages and diseases among others. Explore PCAWG data in EBI Expression Atlas at https://goo.gl/TsIYE5.



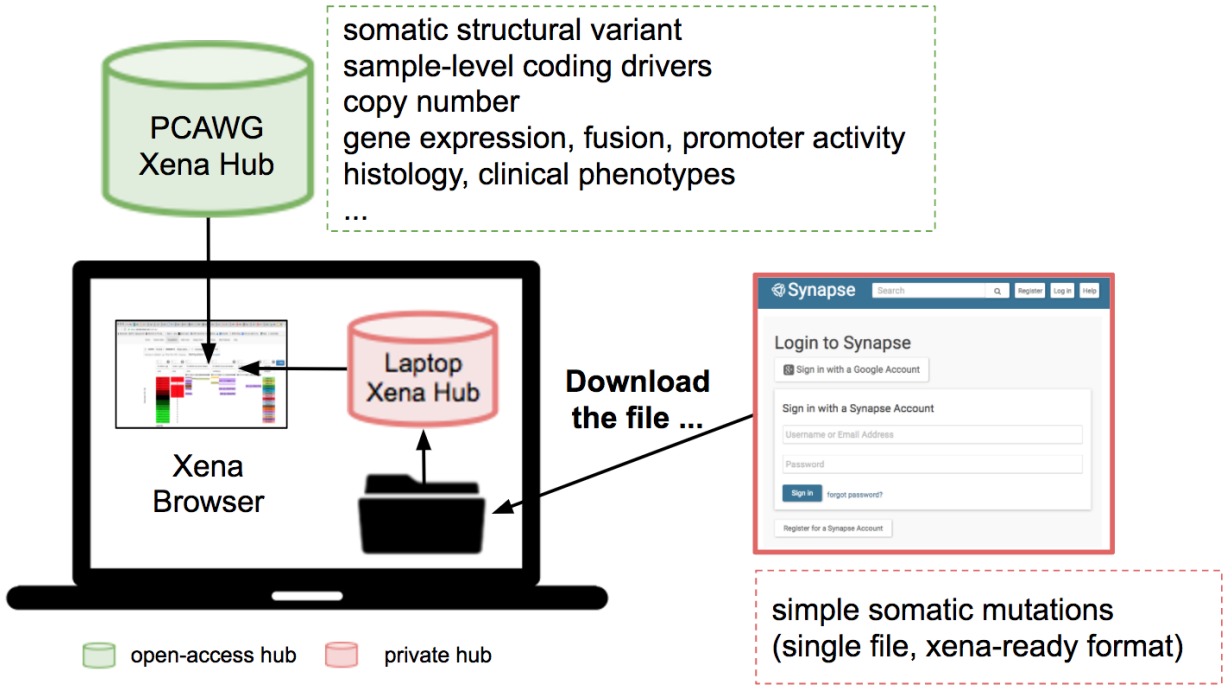
CNIO PCAWG-Scout

PCAWG-Scout is an analysis platform to visualize and explore the PCAWG data. It consists of a portal that presents the original omics data and sample annotation along with the results from different analysis working-groups, and which servers as interface to an on-demand analysis infrastructure to help the user find interesting stories and follow them across different analysis and visualization tools. Explore PCAWG data in CNIO PCAWG-Scout at http://pcawgscout.bioinfo.cnio.es.



Supplementary Figure 2. View protected data using a local Xena Hub.

To view the controlled-access non-coding simple mutations, login to Synapse and down the file containing these mutations (https://www.synapse.org/#!Synapse:syn7122445). Only authorized users can download this protected data. This file is pre-formatted to be imported directly it into a local xena hub (https://genome-cancer.ucsc.edu/download/public/get-xena/index.html) on a user's laptop. Once it is loaded, the UCSC Xena Browser will connect to both the local hub and the public PCAWG hub concurrently, while still keeping the protected data private. This allows users to visualize the whole genome simple mutation data alongside the open-access PCAWG data. More information about using Xena private data hubs can be found at http://xena.ucsc.edu/private-hubs/ .

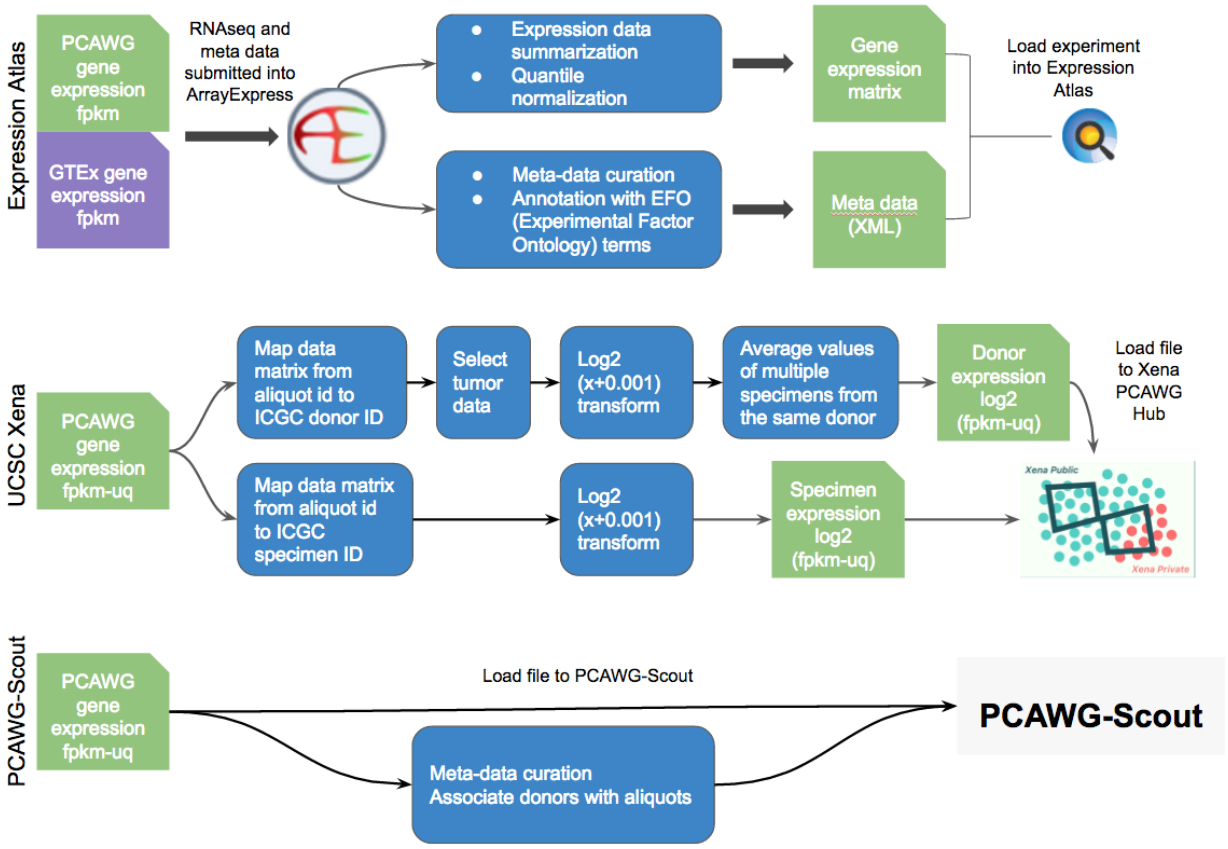


Supplementary Figure 3. RNAseq gene expression data wrangling (processing and curation) by PCAWG online resources.

RNAseq data are visualized by Expression Atlas, UCSC Xena and PCAWG-Scout. Each resource started with the same primary results generated by the analysis working group, and subsequently further processed, curated and refined to meet each resources' quality-control and visualization requirements. The secondarily processed data is displayed on the web.

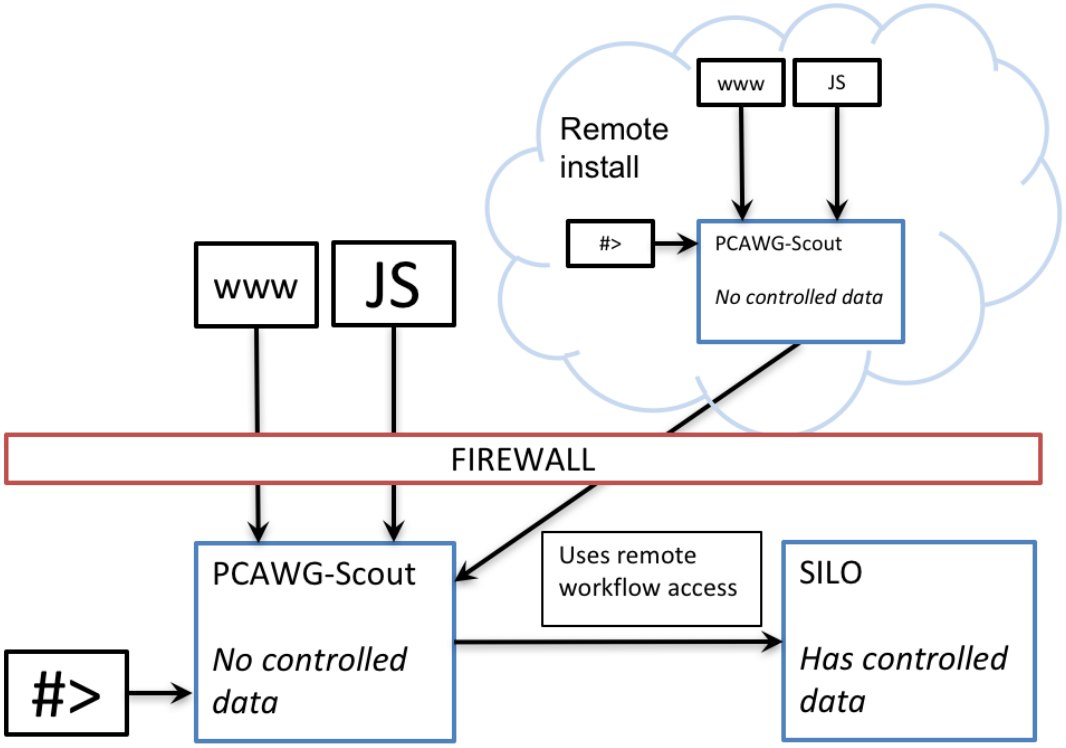
Gene expression preprocessing carried out by Xena and Expression Atlas are illustrated in the figure. GTEx data displayed in Expression Atlas were re-aligned and gene expression re-called using the PCAWG RNA-seq gene level quantification SOP [cite PCAWG-3 paper].

PCAWG-Scout does not perform general pre-processing of the gene expression data. Gene expression data are specifically processed for each visualization or analysis in PCAWG-Scout. For differential expression the values are log2 transformed, with 'no expression' replaced by the smallest number found in the matrix. For differential expression and for expression boxplots on a single gene, only tumor samples are considered and all possible samples for every donor in the group are shown together. When using a color gradient to represent expression of a gene in a donor, all tumor samples for that donor are averaged and the expression is compared with the rest of values for the other tumor samples in the cohort; the rank of that value in the list for all samples in the cohort is used to define the gradient.


Expression Atlas
PCAWG gene expression fpkm
GTEx gene expression fpkm
RNAseq and meta data submitted into ArrayExpress
Expression data summarization
Quantile normalization
Meta-data curation
Annotation with EFO (Experimental Factor Ontology) terms
Gene expression matrix
Meta data (XML)
Load experiment into Expression Atlas
UCSC Xena
PCAWG gene expression fpkm-uq
Map data matrix from aliquot id to ICGC donor ID
Select tumor data
Log2 (x+0.001) transform
Average values of multiple specimens from the same donor
Donor expression log2 (fpkm-uq)
Load file to Xena PCAWG Hub
Map data matrix from aliquot id to ICGC specimen ID
Log2 (x+0.001) transform
Specimen expression log2 (fpkm-uq)
Xena Public
Xena Private
PCAWG-Scout
PCAWG gene expression fpkm-uq
Load file to PCAWG-Scout
Meta-data curation
Associate donors with aliquots
PCAWG-Scout

Supplementary Figure 4. Secured PCAWG-Scout installation to support controlled-access data.

To protect controlled access data, in particular genomic mutations, the PCAWG-Scout is configured so that controlled data is only on a siloed machine behind a firewall. Approved analyses that do not compromise the security of the data are made available by configuring a remote workflow access file on the Rbbt installation. These approved analyses can be accessed via web-browser, javascript plotting utilities, or command-line tools. The PCAWG-Scout machine does not hold the controlled access data and the silo is not directly accessible from outside, keeping the data secure. Any remote installation of the PCAWG-Scout can request these analyses, which will in turn relays them to the silo. This enables this system to be extended by the general research community without requiring all researchers to have access rights to the controlled data.



Supplementary Figure 5. PanDrugs clinical recommendations.

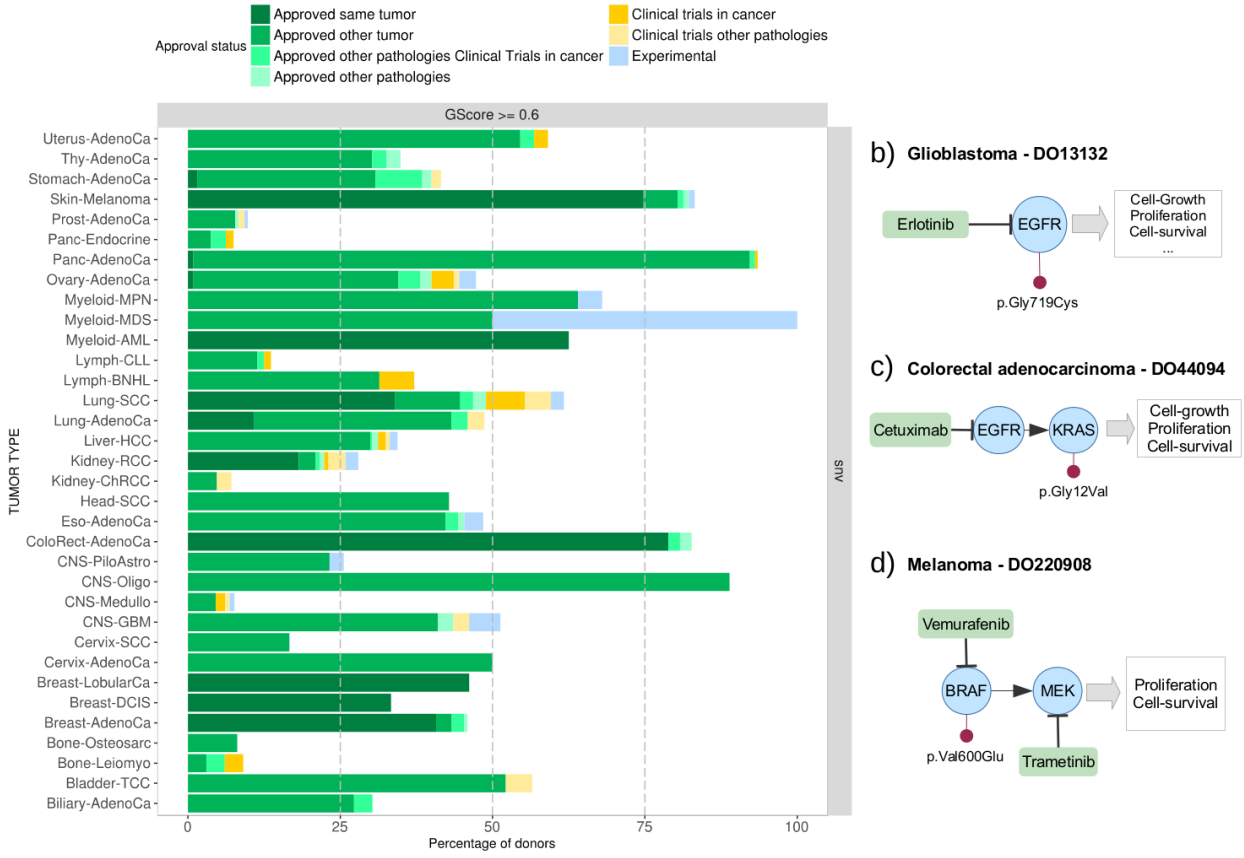
PanDrugs is a web-based tool (http://pandrugs.bioinfo.cnio.es/) to guide the selection of therapies from the results of genome-wide studies in cancers. It allows the identification of actionable molecular alterations and the prioritization of drugs by calculating gene-drug scores (GScore and DScore respectively) [1]. These scores take into account: i) the relevance in cancer of the affected gene and of the concrete variant; ii) the target pathway context; iii) the drug approval status (FDA, clinical trial or experimental small molecule inhibitors); and iv) manually-curated pharmacological information retrieved from the literature. PanDrugs GScore measures the biological relevance in cancer of the gene affected, and the functional impact, and clinical actionability of the specific mutation integrating evidences from public resources. Its DScore measures the suitability of the drug according to the genomic profile. Together they combine the biological and clinical relevance of the genes and their susceptibility to be targeted, reflecting the strength of the evidence of the gene-drug association, and can be used to assist in clinical decision making. Additionally, it incorporates manually curated information about the drug approval status and its usage in cancer therapies or clinical studies in this field.
The current version of PanDrugs integrates data from several sources: DGIdb [2], the tumor alterations relevant for genomics-driven therapy (TARGET) database [3], the Cancer Therapeutics Response Portal [4], the Genomics of Drug Sensitivity in Cancer (GDSC) [5], and additional information about monoclonal antibodies. PanDrugs supports more than 50,000 drug-target associations obtained from around 6000 genes and 11000 unique compounds.

[1] Piñeiro E. et al. PanDrugs: Prioritizing drug treatment in cancer according to individual genomic data (BioRxiv preprint)
[2] Wagner AH. et al. (2016) Nucleic Acids Res. Jan 4;44(D1):D1036-44
[3] Van Allen EM. et al. (2014) Nat Med. Jun;20(6):682-8
[4] Basu A. et al. (2013) Cell. Aug 29;154(5):1151-61
[5] Iorio F. et al. (2016) Cell. Jul 28;166(3):740-54

a) Overview of the drug assignment distribution for 2495 donors from 37 different tumor types. Each bar represents the percentage of patients with a suggested therapy in each tumor type based on detected simple somatic mutations. Different colors correspond to different approval status for the drugs as indicated in the legend. Only high impact alterations, those with a PanDrugs GScore greater or equal to 0.6, are considered. **b**) Example of a therapy suggestion based on evidence from an affected gene. Donor DO13132 with Glioblastoma (CNS-GBM) has a missense mutation in *EGFR* gene (p.Gly719Cys), which leads to carcinogenic processes of cell growth and proliferation. This mutation confers sensitivity to the EGFR inhibitors such as Erlotinib, one of the proposed therapies. **c**) Example of a therapy suggestion against the use of a conventional therapy based on evidence from an affected gene. EGFR inhibitor Cetuximab is a standard therapy for the treatment of colorectal cancer, but *KRAS* mutations have shown to be a predictor of resistance to cetuximab therapy. In the DO44094 case, the missense p.Gly12Val mutation indicates resistance to Cetuximab. **d**) Example of an indirect therapy suggestion based on evidence from a pathway relationship. DO220908 with melanoma has the

p.Val600Glu alteration in *BRAF*. This alteration suggests the administration of BRAF inhibitors such as Vemurafenib, but also, MEK inhibitors, as for example, Trametinib, according to the downstream position of these gene in relation to BRAF in the MAPK signaling pathway.

a) Percentage of donors with treatments suggested by PanDrugs



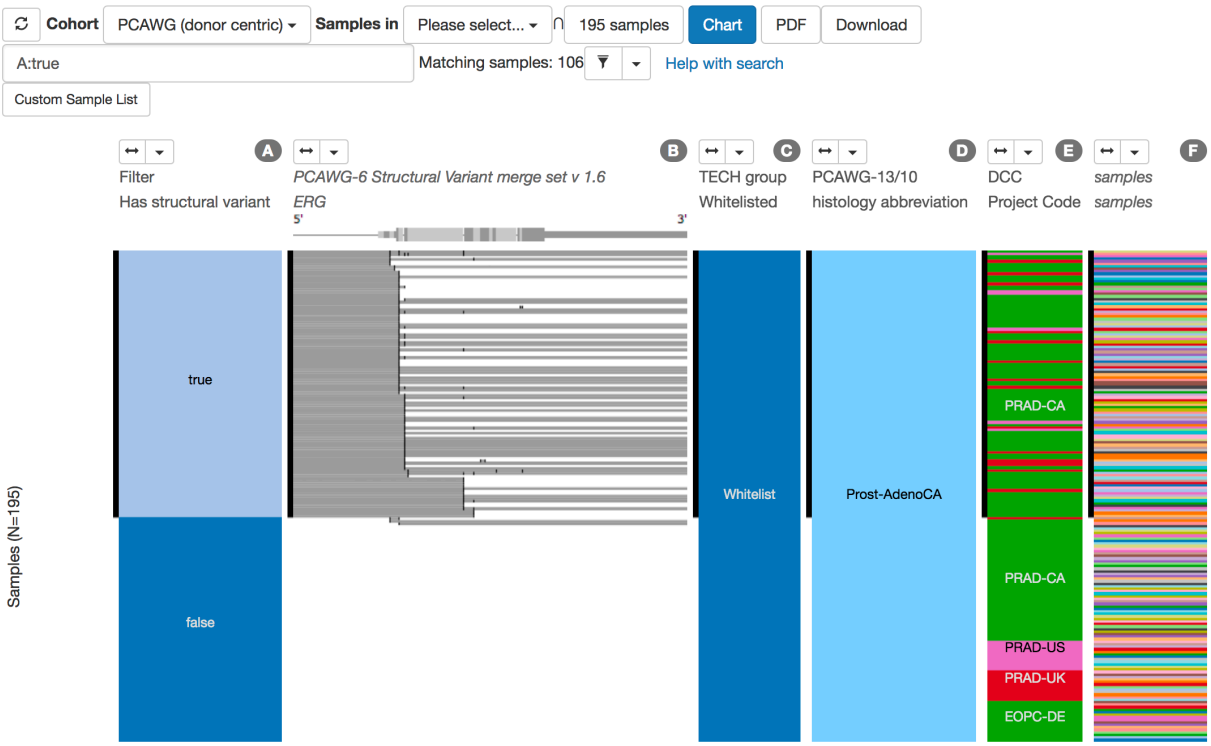
b) **Glioblastoma - DO13132**

c) **Colorectal adenocarcinoma - DO44094**

d) **Melanoma - DO220908**

Supplementary Figure 6. Prostate adenocarcinoma ERG structural variants

Xena Visual Spreadsheet shows 106 of the 195 prostate adenocarcinoma tumors have the ERG fusion that joins an external piece of DNA from the 5' direction, replacing the the 5' end of ERG with another piece of DNA (promoter fusion). These 195 prostate tumors come from four different projects: PRAD-CA, PRAD-US, PRAD-UK and EOPC-DE. 89 out of the 195 prostate tumors (46%) do not have this type of fusion detected. Three samples have structural variants detected in ERG, but they are not the 5' jointing type (DO52498, DO50430, and DO51087). The Xena view in the figure illustrating all ERG structural variants in PCAWG prostate tumors can be accessed here: https://xenabrowser.net/heatmap/?bookmark=f69523bca67afd99d18c19a74161be8d. For the structural variants (column B), the grey-colored lines represent the external DNA that are fused to ERG, and the short black ticks mark the breakpoints in ERG. Breakpoints are clustered at the beginning of the coding regions.



Supplementary Table 1. List of PCAWG online data resources.

Resource Home Page	Resource PCAWG Landing Page	Functionality offered
ICGC Data Portal https://dcc.icgc.org	http://docs.icgc.org/pcawg	Data Portal and PCAWG project Landing page Search and download PCAWG BAMs, VCFs, primary working groups results, exploration of PCAWG consensus somatic mutations integrated with clinical data and rich annotations
PCAWG-Scout http://pcawgscout.bioinfo.cnio.es/	http://pcawgscout.bioinfo.cnio.es/	On-demand analysis infrastructure Visualization Public views of protected somatic mutation data
UCSC Xena http://xena.ucsc.edu	https://pcawg.xenahubs.net	Visualization of all primary PCAWG analysis working group results Integrate investigator generated private data with PCAWG data
Expression Atlas http://www.ebi.ac.uk/gxa	http://www.ebi.ac.uk/gxa/experiments/E-MTAB-5200?accessKey=1522478c-1bd0-4863-848e-d15e86774418	Visualization of gene expression data Visualization of tumour (e.g. PCAWG) gene expression, or tumour and healthy tissue (e.g. GTEX) expression comparisons.

Supplementary Table 2. List of PCAWG primary results supported by online visualization resources.

Table of primary results generated by PCAWG analysis working groups available for visualization by UCSC Xena, Expression Atlas and PCAWG-Scout. Each primary result is referenced by corresponding synapse IDs. Synapse folder ID is the identifier for the synapse landing page for each type of primary results. The landing page typically includes a summary written by the analysis working group to briefly describe the bioinformatics methods used and a list of results generated. Because there are often multiple versions of the same results files (such as fpkm vs fpkm-uq gene expression estimations, or simple mutations from all specimens or aggregated by donors), synapse identifiers in the remaining columns point to the actual data file ingested by each online resource. The data snapshot was taken as of Feb 10, 2017.

Data	Synapse page ID	UCSC Xena	Expression Atlas	PCAWG-Scout
Consensus SNVs and indels	syn7118450	syn7364923 syn7364924		syn7364923
Consensus SVs	syn5964535	syn7596712		syn7596712
Consensus copy number	syn8042880	syn8042988		syn8042992
Gene expression	syn3104297	syn5553991	syn5553983 syn5553985	syn5553991
GTEx gene expression derived using the PCAWG RNA-seq SOP	syn8105922		syn8105922	
RNAseq gene fusion	syn7221157	syn7221157		
RNAseq alternative promoter usage	syn3354819	syn7247455		

small RNA-Seq (miRNA) analyses	syn5842981	syn5878064 syn5878067		
Patient-centric driver catalogue - Coding	syn7250536	syn7328242		syn7328242
Integrated driver calls	syn7359546			syn8035740
APOBEC mutagenesis analysis	syn7437205	syn7511424		
Tumour subtype and histology information	syn2364731	syn7253569	syn7253569	syn7253569
Donor clinical data	syn2364731	syn7772065		syn7772065

Supplementary Table 3. PCAWG data available through UCSC Xena visualization.

Primary PCAWG analysis working group results	Xena Hub
Consensus simple somatic mutations	Local hub
Consensus simple somatic mutations - coding	PCAWG hub
Patient-centric driver catalogue - coding	PCAWG hub
Consensus somatic structural variants	PCAWG hub
Consensus copy number	PCAWG hub
RNAseq gene expression	PCAWG hub
RNAseq alternative promoter usage	PCAWG hub
RNAseq gene fusion	PCAWG hub
miRNA expression	PCAWG hub
Tumor purity and ploidy	PCAWG hub
APOBEC mutagenesis analysis	PCAWG hub

Specimen histopathology, molecular subtype, donor clinical data	PCAWG hub
Included, excluded and grey-listed donors and samples	PCAWG hub

Supplementary Table 4. Code availability.

Resource	Open source code availability
ICGC Data Portal	https://github.com/icgc-dcc/dcc-portal
PCAWG-Scout	http://mikisvaz.github.io/rbbt/; https://github.com/Rbbt-Workflows; https://github.com/Rbbt-Apps/PCAWGScout
UCSC Xena Browser	https://github.com/ucscXena/ucsc-xena-client
Expression Atlas	https://github.com/gxa/atlas

Supplementary Table 5. Embeddable javascript modules.

Javascript Module	Utility	Open source code availability
OncoGrid	Generate OncoGrids and related tracks	https://github.com/oncojs/oncogrid
Xena Visual Spreadsheet	Generate visual spreadsheet	https://github.com/ucscXena/ucsc-xena-client
Kaplan-Meier	Kaplan-meier estimator and log-rank test	https://github.com/ucscXena/kaplan-meier
static-interval-tree	Fast overlapping interval queries in javascript	https://github.com/ucscXena/static-interval-tree
Expression Atlas Widget: Heatmap & Anatomogram	View tissue-specific results on a heatmap and human figure	https://github.com/gxa/atlas-heatmap