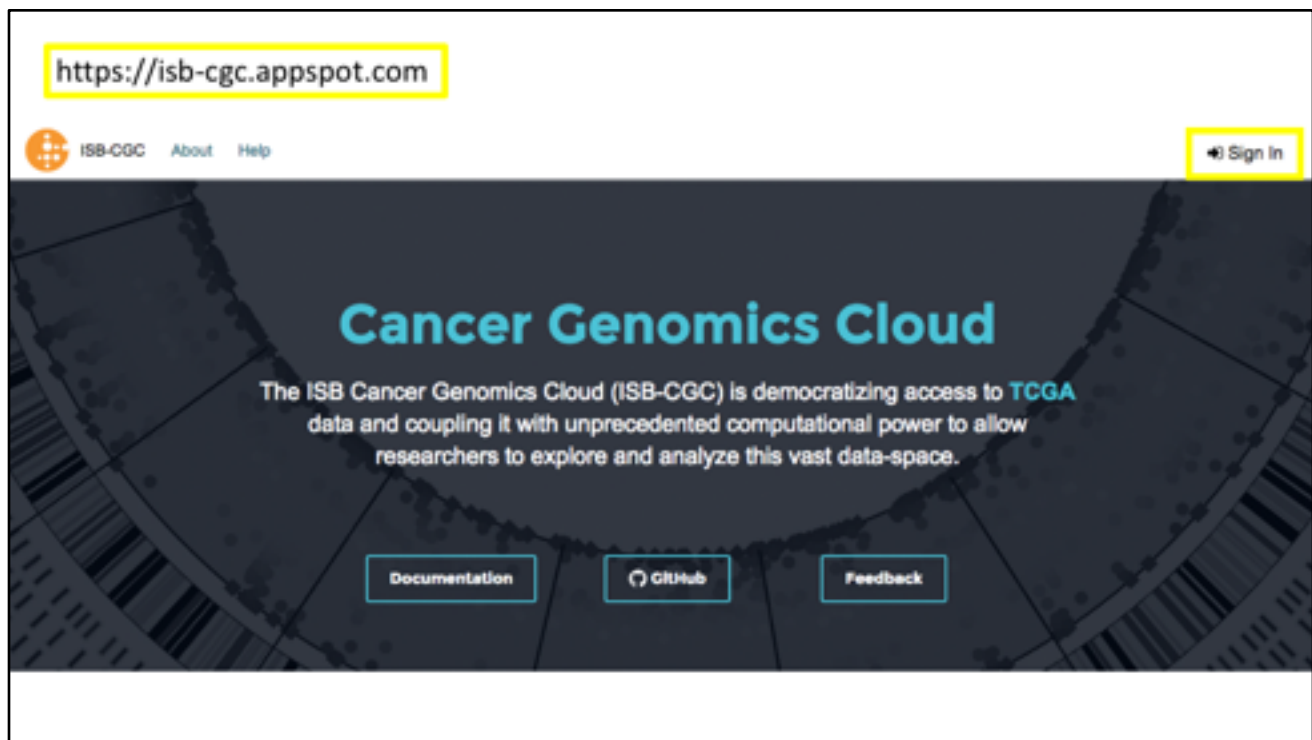


An Introduction to the ISB-CGC Web App

brought to you by

The ISB Cancer Genomics Cloud





- This is our main landing page.
- There are a few links here that you can use to get to documentation, code, and send us feedback.
- You may only log in using a Google managed identity by clicking the Sign In button.

1. Log into the system:

Your Dashboard

Saved Workbooks (0)

Workbooks store the Analyses you create -- and their related data.

[+ Create A New Workbook](#)

Saved Cohorts (0)

You don't have any saved Cohorts.

[Create Cohort](#)

Gene Favorites (0)

You don't have any saved Gene Favorites.

[Create Gene Favorites](#)

Variable Favorites (0)

You don't have any saved Variable Favorites.

[Create Variable Favorites](#)

- After logging in, you are taken to the dashboard.
- This is where you can view an overview of the different workbooks and cohorts you create.
- Workbooks contain worksheets, where you can create analyses.
- Gene and Variable favorites is where you can define lists of interest to yourself.
- On top of this, there is a Menu button next to your username that you can use to easily jump from page to page.

2. Click "Create Cohort"

Your Dashboard

Saved Workbooks (0)

Workbooks store the Analyses you create -- and their related data.

[+ Create A New Workbook](#)

Saved Cohorts (0)

You don't have any saved Cohorts.

[Create Cohort](#)

Gene Favorites (0)

You don't have any saved Gene Favorites.

[Create Gene Favorites](#)

Variable Favorites (0)

You don't have any saved Variable Favorites.

[Create Variable Favorites](#)

Dashboard > Cohorts
Create Cohort

DONOR

- PUBLIC PROJECTS
 - ☒ TCGA (2066)
 - ☐ GSE (179)
- PUBLIC STUDIES
- VITAL STATUS
- GENDER
- AGE AT DIAGNOSIS
- SAMPLE TYPE
- TUMOR TISSUE SITE
- HISTOLOGICAL TYPE
- PRIOR DIAGNOSIS
- PATHOLOGIC STAGE
- TUMOR STATUS
- NEW TUMOR EVENT AFTER INITIAL TREATMENT
- HISTOLOGICAL GRADE
- RESIDUAL TUMOR
- TOBACCO SMOKING HISTORY
- ICD-10
- ICD-O-3 SITE

DATA TYPE

Selected Filters
Clear All

Project: TCGA ✕

Details

Total Number of Samples:	2066	Total Number of Participants:	11311
--------------------------	------	-------------------------------	-------

Clinical Features

Study

Vital Status

Sample Type

Tumor Tissue Site

[Show More](#)

Public Data Availability

SNP/CN	DNAseq	DNAmeth	mRNA	microRNA	Protein

- 5

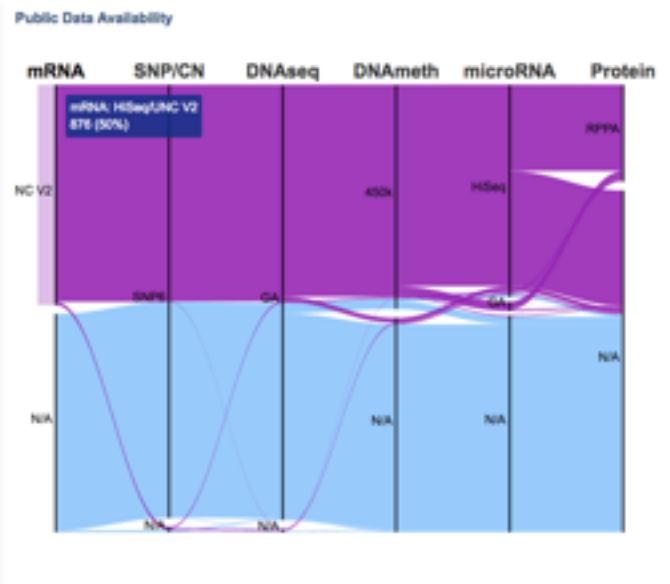
4. Create TCGA Head and Neck (HNSC), and Cervical (CESC) Cohort

The image displays three sequential screenshots of the TCGA data selection interface, illustrating the steps to create a cohort for Head and Neck (HNSC) and Cervical (CESC) cancer.

- First Screenshot:** The 'DONOR' tab is selected. Under 'PUBLIC STUDIES', the 'TCGA (20986)' project is selected (indicated by a blue dot), and the 'CCLE (1796)' project is unselected. The 'PUBLIC STUDIES' section is highlighted with a yellow box.
- Second Screenshot:** The 'DATA TYPE' tab is selected. Under 'PUBLIC STUDIES', the 'HNSC (1196)' data type is selected (indicated by a blue dot), and the 'CCLE (1796)' data type is unselected. The 'HNSC (1196)' section is highlighted with a yellow box.
- Third Screenshot:** The 'DATA TYPE' tab is selected. Under 'PUBLIC STUDIES', the 'CESC (875)' data type is selected (indicated by a blue dot), and the 'HNSC (1196)' data type is unselected. The 'CESC (875)' section is highlighted with a yellow box.

- For the purposes of our analysis, we will create a cohort comprised of all TCGA Head and Neck and Cervical samples.
- To do this we select those from the Public Studies.
- It is important to note that if we had not selected the TCGA Project, our cohort could include samples that are also from the CCLE Project.

5. Let's look at data availability for this cohort

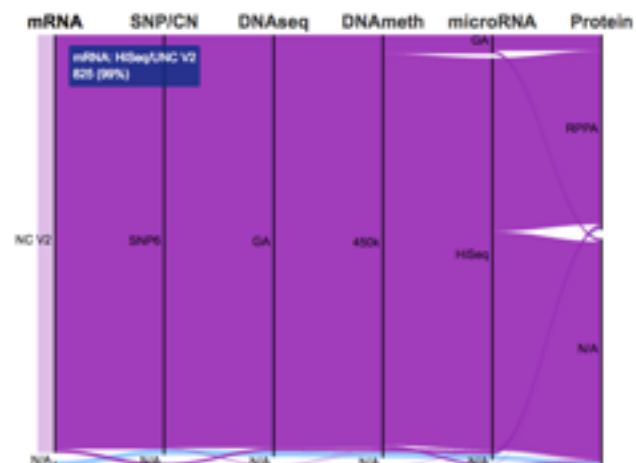


- This is called a parallel sets graph. It shows the distribution of data for the samples selected.
- 50% of our participants have HiSeq/UNC V2 gene expression data available
- Of those 876 samples, we can see that a large portion of them have SNP6 data, and a small sliver do not.
- Of the samples that have both HiSeq/UNC V2 and SNP6 data, another large portion also have DNaseq: GA data.

- The data availability graph can be re-ordered based on what you're most interested in. Here, we use gene expression data as our main focus.

6. Select the Sample Type 'Primary tumor Tissue'

Public Data Availability



- Notice that now 99% of our samples have HiSeq/UNC V2 gene expression data.

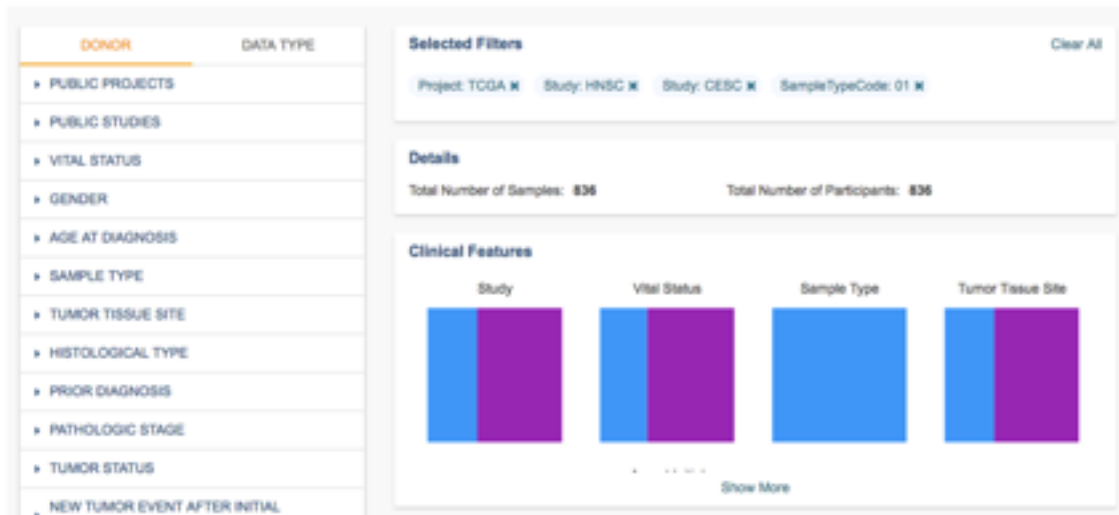
- After selecting only Primary tumor Tissue, we can see that most of our samples have gene expression data.

7. The resulting cohort

Your Dashboard > Cohorts >

Create Cohort

Save As New Cohort



- The resulting cohort should look something like this.
- You can see the filters used to create the cohort.
- There are 836 samples and the same number of participants.
- If you click on the 'Show More' in the Clinical Features, you will see a breakdown of the age and gender of your cohort.

8. Save the cohort and provide it a name: TCGA Head and Neck, and Cervical

Your Dashboard > Cohorts >

Create Cohort

Save As New Cohort

DONOR	DATA TYPE
» PUBLIC PROJECTS	
» PUBLIC STUDIES	
» VITAL STATUS	
» GENDER	
» AGE AT DIAGNOSIS	
» SAMPLE TYPE	
» TUMOR TISSUE SITE	
» HISTOLOGICAL TYPE	
» PRIOR DIAGNOSIS	
» PATHOLOGIC STAGE	
» TUMOR STATUS	
» NEW TUMOR EVENT AFTER INITIAL	

Selected Filters

Clear All

Project: TCGA X Study: HNSC X Study: CESC X SampleTypeCode: 01 X

Details

Total Number of Samples: 836 Total Number of Participants: 836

Clinical Features

Study Vital Status Sample Type Tumor Tissue Site

Show More

- We will now save the cohort with this name: TCGA Head and Neck, and Cervical

9. Cohort Listing Page

Your Dashboard >

Cohorts

+ Create New Cohort

SAVED COHORTS

PUBLIC COHORTS

New Workbook

Delete

Set Operations

Share

<input type="checkbox"/>	Cohort Name	# Samples	# Patients	Ownership	Shared With	Last Modified	⌵
<input type="checkbox"/>	TCGA Head and Neck, and Cervical	836	836	You	(0)	05/16/2016 9:32 a.m.	
<input type="checkbox"/>	TCGA HNSC CESC	1744	836	You	(0)	05/16/2016 12:13 p.m.	
<input type="checkbox"/>	TCGA HNSC	1123	528	You	(0)	05/16/2016 12:12 p.m.	
<input type="checkbox"/>	TCGA CESC	621	308	You	(0)	05/12/2016 12:02 p.m.	
<input type="checkbox"/>	DNA Meth	213	214	You	(0)	04/12/2016 2:30 p.m.	
<input type="checkbox"/>	GBM	1174	605	You	(0)	03/14/2016 11:10 a.m.	

- This is where you can see all of the cohorts you've created and that have been shared with you.
- You'll notice that you also have access to Public Cohorts. These are cohorts that we've created for you. So far it's just one, but we plan on adding more.
- Another way that we could have created our cohort is by taking the union of two previously created cohorts. In this example, you can see that there is already a TCGA HNSC and TCGA CESC cohort. I could select those and click the Set Operations button. We currently support Unions, Intersects, and Set Complements.
- To start an analysis, we're going to select our cohort and click the New Workbook button. We're going to use this cohort and explore differential gene expression conditional on HPV Status.

10. New Workbook

[Your Dashboard](#) > [Saved Workbooks](#) >

Untitled Workbook

This is a workbook created with cohorts added to the first worksheet. Click Edit Details to change your workbook title and description.

[Edit Details](#) [Duplicate](#) [Delete](#) [Share](#) Shared With (0)

Worksheet 1 +

Comments (0)

Source Data

Genes +

Variables +

Cohorts +

TCGA Head and Neck, and Cervical

Analysis Type

-- select an analysis --

Edit Analysis Settings

To Complete this Analysis:

- You must select an Analysis Type (above)
- You must select Genes or Variables (or, optionally, both)
- You must select a Cohort

Resubmit Plot

- When you create a new workbook, it is automatically populated with one worksheet.
- A worksheet is comprised of different data sources that you will use in your analysis. You can see that the Cohort we selected is already available.
- Let's first edit some details of our workbook by giving it a more meaningful name and then a short description.

11. Add Variables to your worksheet

[Your Dashboard](#) > [Saved Worksheets](#) >

Untitled Workbook

This is a workbook created with cohorts added to the first worksheet. Click [Edit Details](#) to change your workbook title and description.

[Edit Details](#) [Duplicate](#) [Delete](#) | [Share](#) Shared With (2)

Worksheet 1 +

Comments (0)

Source Data

Genes +

Variables +

Cohorts +

TCGA Head and Neck, and Cervical

Analysis Type

-- select an analysis --

✦ Edit Analysis Settings

To Complete this Analysis:

- You must select an Analysis Type (above)
- You must select Genes or Variables (or, optionally, both)
- You must select a Cohort

Resubmit Plot

12. Creating a new Variable List

Your Dashboard > Saved Workbooks > HPV Workbook > Saved Variable Favorites >

Data Source | Variables

[Apply To Worksheet](#) [Back To Workbook](#)

Name of Favorite (Required)

COMMON FAVORITES (8) CLINICAL MIRNA

- ☐ VITAL STATUS
- ☐ GENDER
- ☐ AGE AT DIAGNOSIS
- ☐ TUMOR TISSUE SITE
- ☐ HISTOLOGICAL TYPE
- ☐ PRIOR DIAGNOSIS
- ☐ TUMOR STATUS
- ☐ NEW TUMOR EVENT AFTER INITIAL TREATMENT
- ☐ HISTOLOGICAL GRADE
- ☐ RESIDUAL TUMOR
- ☐ TOBACCO SMOKING HISTORY
- ☐ ICD-10
- ☐ ICD-O-3 SITE
- ☐ ICD-O-3 HISTOLOGY

Selected Variables [Clear All](#)

Select your favorite variables from the left panel.

- If you don't already have variable lists created, you will be taken here. If you do, then you will be taken to your list of previously created variable lists. To get to this page, click the Apply New Variable List button.
- The idea behind this concept is for you to be able to create a list of variables you might use in your analysis and save it all together. It will also allow you to reuse that list in other analyses.
- Here, you can select variables that are **not** gene specific, so mainly clinical and miRNA.

13. Provide a name and select the following variables from the Common tab.

The screenshot shows a web interface for selecting variables. At the top, there is a text input field labeled "Name of Favorite (Required)" with the text "HPV Variables" entered. Below this is a tabbed interface with four tabs: "COMMON", "FAVORITES (8)", "CLINICAL", and "MRNA". The "COMMON" tab is currently selected. Under the "COMMON" tab, there is a list of variables, each with a selection icon (a blue square with a white 'x' or a radio button). The selected variables are: VITAL STATUS, GENDER, AGE AT DIAGNOSIS, TUMOR TISSUE SITE, HISTOLOGICAL TYPE, PRIOR DIAGNOSIS, TUMOR STATUS, TOBACCO SMOKING HISTORY, ICD-10, ICD-O-3 SITE, and ICD-O-3 HISTOLOGY. To the right of the variable list is a panel titled "Selected Variables" with a "Clear All" link. This panel contains the text "Select your favorite variables from the left panel." and a list of the selected variables: Vital Status, Gender, Age at Diagnosis, Tumor Tissue Site, Histological Type, Prior Diagnosis, Tumor Status, and Tobacco Smoking History.

- We provide a name for our variable list: HPV Variables
- And select the following variables on the common tab:
 - Vital Status
 - Gender
 - Age at Diagnosis
 - Tumor Tissue Site
 - Histological Type
 - Prior Diagnosis
 - Tumor Status
 - Tobacco Smoking History
- You'll notice that they will appear in the Selected Variables panel.

14. Select HPV Calls, HPV Status, and Study from the Clinical tab

The screenshot shows a web interface for selecting variables. At the top, there is a text input field labeled "Name of Favorite (Required)" containing the text "HPV Variables". Below this is a tabbed interface with four tabs: "COMMON", "FAVORITES (8)", "CLINICAL" (which is highlighted in orange), and "MPNA". Under the "CLINICAL" tab, there is a "Feature Search" dropdown menu. This dropdown is highlighted with a yellow rectangular box, and a blue arrow points from the text below to it. The dropdown menu currently shows the word "Study". To the right of the "Feature Search" dropdown is a "Selected Variables" panel. This panel has a "Clear All" link at the top right. Below the link, it says "Select your favorite variables from the left panel." and lists several variables, each with a small 'x' icon to its right: "Vital Status", "Gender", "Age at Diagnosis", "Tumor Tissue Site", "Histological Type", "Prior Diagnosis", "Tumor Status", "Tobacco Smoking History", "Hpv Calls", "Hpv Status", and "Study".

This is an autocomplete box, so try typing in 'hpv' to get the HPV specific variables

15. Save the list by clicking the "Apply to Worksheet" button

- We also want some less common clinical variables, so we move on to the Clinical tab.
- Here we can start typing in the variable we're interested in. In our case it's 'hpv'
- To get the Study variable, try using just part of the work like 'tud'
- We hit save and are brought back to the worksheet.

16. Add Genes to your worksheet

Your Dashboard > Saved Workbooks >

HPV Workbook

Edit Details

Duplicate

Delete

Share

Shared With (2)

Worksheet 1



Comments (0)

Source Data

Genes

Variables

Vital Status

Gender

Age at Diagnosis

Tumor Tissue Site

Histological Type

Prior Diagnosis

Analysis Type

-- select an analysis --

Edit Analysis Settings

To Complete this Analysis:

- You must select an Analysis Type (above)
- You must select Genes or Variables (or, optionally, both)
- You must select a Cohorts

Resubmit Plot

17. Create a gene list for your HPV analysis

Your Dashboard > Saved Workbooks > HPV Workbook > Saved Gene Favorites >

Create Gene List

Name of Favorite (Required)

HPV Genes

Selected Genes (required) Clear All Upload Genes List

PVT1 x RAD51L1 x TMPRSS3 x ERBB2 x FN1 x SERPINB11 x Enter your favorite gene's name

Apply To Worksheet Cancel View Gene Identifiers

This is an autocomplete box, so try typing in 'RAD51'

18. When complete, click 'Apply to Worksheet' to save and return to your workbook

- Similarly to variables, if you have gene lists created, you will be taken to a listing of your gene lists.
- This page uses an autocomplete to help you find the genes you're looking for. Try typing in some of your favorite genes to see if they are in our system. If you're unsure of what your gene might be called, you can use the View Gene Identifiers to help.
- We are going to use this list of genes:
 - PVT1
 - RAD51L1
 - TMPRSS3
 - ERBB2
 - FN1
 - SERPINB11
- We provide a name, and click the Apply To Worksheet button.

19. Creating a Violin Plot comparing HPV Status VS PVT1 Gene Expression

The screenshot shows a web-based data analysis tool interface. On the left, there's a 'Source Data' panel with 'Genes' (ERBB2, FN1, PVT1, SCRN4B11, RAD51L1, TMPOSS) and 'Variables' (Vital Status, Gender, Age at Diagnosis, Tumor Tissue Site, Histological Type, Prior Diagnosis, Tobacco Smoking History, Tumor Status, Residual Tumor, Hpa Cells, Hpa Status, Study). Below these is a 'Cohorts' section with 'TCGA Head and Neck, and Cervical'. The main panel is titled 'Analysis Type' and has 'Violin Plot' selected. Below this, it says 'To Complete this Analysis: You must select an Analysis Type (above); You must select Genes or Variables (or, optionally, both); You must select a Cohort'. A 'Recompute Plot' button is present. On the right, the 'Plot Settings' panel is visible. The 'X Axis Variable' dropdown is highlighted with a yellow box, and a blue arrow points to it with the text 'Select 'HPV Status''. Below it are 'Y Axis Variable' and 'Color By Feature' dropdowns, all currently showing '-- select a variable --'. At the bottom of the 'Plot Settings' panel, there's a 'Cohorts' section with a radio button for 'TCGA Head and Neck, and Cervical' and an 'Update Plot' button.

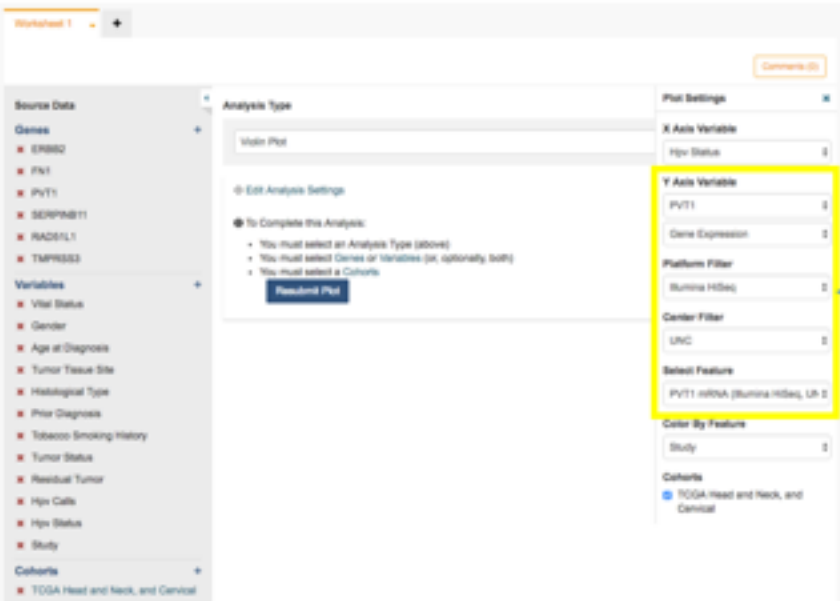
- Now that we have all our data sources ready, we want to start our analysis.
- We provide several different types of analyses (For more information please see our online documentation):
 - Barchart – 1 Categorical variable
 - Histogram – 1 Numerical variable
 - Scatterplot – 2 Numerical variables
 - Violin Plot – 1 Categorical and 1 Numerical variable
 - Cubby Hole Plot – 2 Categorical variables
 - SeqPeek – 1 Gene
- We want to plot HPV Status VS Gene Expressions for PVT1. Since that is a categorical feature versus a numerical feature, we choose a violin plot.
- We first select 'HPV Status' for the X-Axis

19. Creating a Violin Plot comparing HPV Status VS PVT1 Gene Expression

The screenshot shows a software interface for creating a violin plot. On the left, a sidebar lists 'Source Data' categories: Genes (ERBB2, PVT1, SERPINE1, RGS12, TMPSB3), Variables (Vital Status, Gender, Age at Diagnosis, Tumor Tissue Site, Histological Type, Prior Diagnosis, Tobacco Smoking History, Tumor Status, Residual Tumor, Hpa Cells, Hpa Status, Study), and Cohorts (TCGA Head and Neck, and Cervical). The main panel is titled 'Analyze Type' and shows 'Violin Plot' selected. Below this, it says 'To Complete this Analysis: You must select an Analyze Type (above), You must select a Series or Variables (or, optionally, both), and You must select a Cohort'. A 'Replot Plot' button is visible. On the right, the 'Plot Settings' panel shows 'X Axis Variable' set to 'Hpa Status' and 'Y Axis Variable' set to 'PVT1'. A blue arrow points to the 'Y Axis Variable' dropdown with the text 'Select 'PVT1''. Below this, the 'Color By Feature' section has a dropdown set to 'Please select an option'. The 'Cohorts' section has a checkbox for 'TCGA Head and Neck, and Cervical' which is unchecked. An 'Update Plot' button is at the bottom right.

- Then we select PVT1 for our Y-Axis

19. Creating a Violin Plot comparing HPV Status VS RAD51L1 Gene Expression



Source Data

Genes

- ERBB2
- FN1
- PVT1
- SERPINB1
- RAD51L1
- TMPRSS3

Variables

- Vital Status
- Gender
- Age at Diagnosis
- Tumor Tissue Site
- Histological Type
- Prior Diagnosis
- Tobacco Smoking History
- Tumor Status
- Residual Tumor
- Hpa Cells
- Hpa Status
- Study

Cohorts

- TCGA Head and Neck, and Cervical

Analysis Type

Violin Plot

Edit Analysis Settings

To Complete this Analysis:

- You must select an Analysis Type (above)
- You must select Genes or Variables (or, optionally, both)
- You must select a Cohort

Platform Filter

Illumina HiSeq

Center Filter

UNC

Select Feature

PVT1 mRNA (Illumina HiSeq, LP 2)

Center By Feature

Study

Cohorts

- TCGA Head and Neck, and Cervical

- Gene Expression
- Platform: Illumina HiSeq
- Center: UNC
- Feature: PVT1 mRNA (Illumina HiSeq, UNC RSEM)

- You will notice more options appear.
- First we select Gene Expression
- Specify a Platform and Center, then finally we select the actual variable we'd like to plot. Without specifying the platform and filter, we could end up with a lot of potential variables to plot, but can only pick one at a time.

19. Creating a Violin Plot comparing HPV Status VS RAD51L1 Gene Expression

Source Data

Genes

- ERBB2
- FN1
- PVT1
- SERPINE1
- RAD51L1
- TNFRSF5

Variables

- Vital Status
- Gender
- Age at Diagnosis
- Tumor Tissue Site
- Histological Type
- Prior Diagnosis
- Tobacco Smoking History
- Tumor Status
- Residual Tumor
- HPV Cells
- HPV Status
- Study

Cohorts

- TCGA Head and Neck, and Cervical

Analysis Type

Violin Plot

Plot Settings

HPV Status

Y Axis Variable

PVT1

Gene Expression

Platform Filter

Summa HBB

Gender Filter

UNC

Select Feature

PVT1 mRNA (Summa HBB), L1

Color By Feature

Study

Cohorts

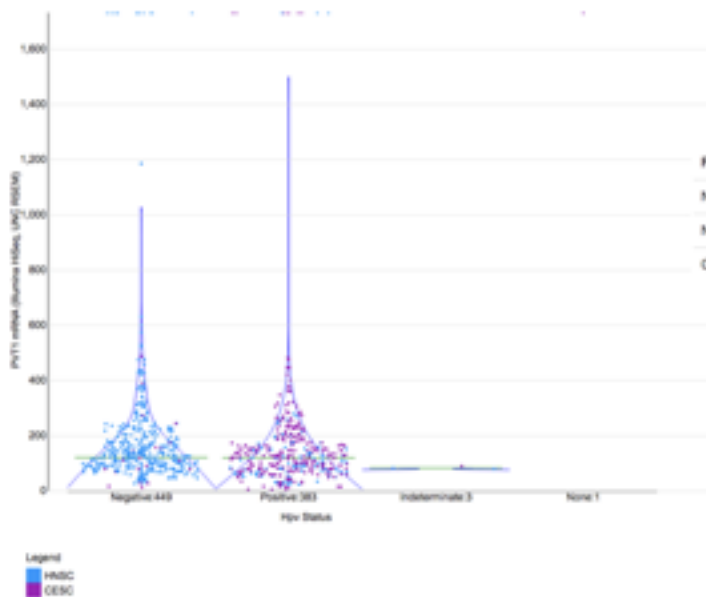
- TCGA Head and Neck, and Cervical

Update Plot

- Color By: Study
- Select Cohort
- Update Plot

- Next we want to color by Study. The violin plot will show each sample as a dot. By adding a color by, we are able to see an extra dimension of data.
- We also select the cohort we're interested in. If you had multiple cohorts in your data sources, you can select more than one.
- And we click the Update Plot button.

19. Creating a Violin Plot comparing HPV Status VS RAD51L1 Gene Expression



- This is the resulting violin plot.
- You can see that there are a lot more CESC samples that are HPV Positive

Feature 1	Feature 2	logp	n
N:GEXP-PVT1:mma_unc illumina_hiseq	C:CLIN:hpr_status	0.752	815
N:GEXP-PVT1:mma_unc illumina_hiseq	C:CLIN:Study	0.378	815
C:CLIN:hpr_status	C:CLIN:Study	105.326	832

- Sample pairwise results for the features selected.

- This is the resulting plot.
- The expression pattern is not significantly different with the HPV status for the samples we've chosen. Let's move to more detailed analysis using R to see if we can pick up any significant patterns of interest.