# An Introduction to the ISB-CGC Web App

brought to you by

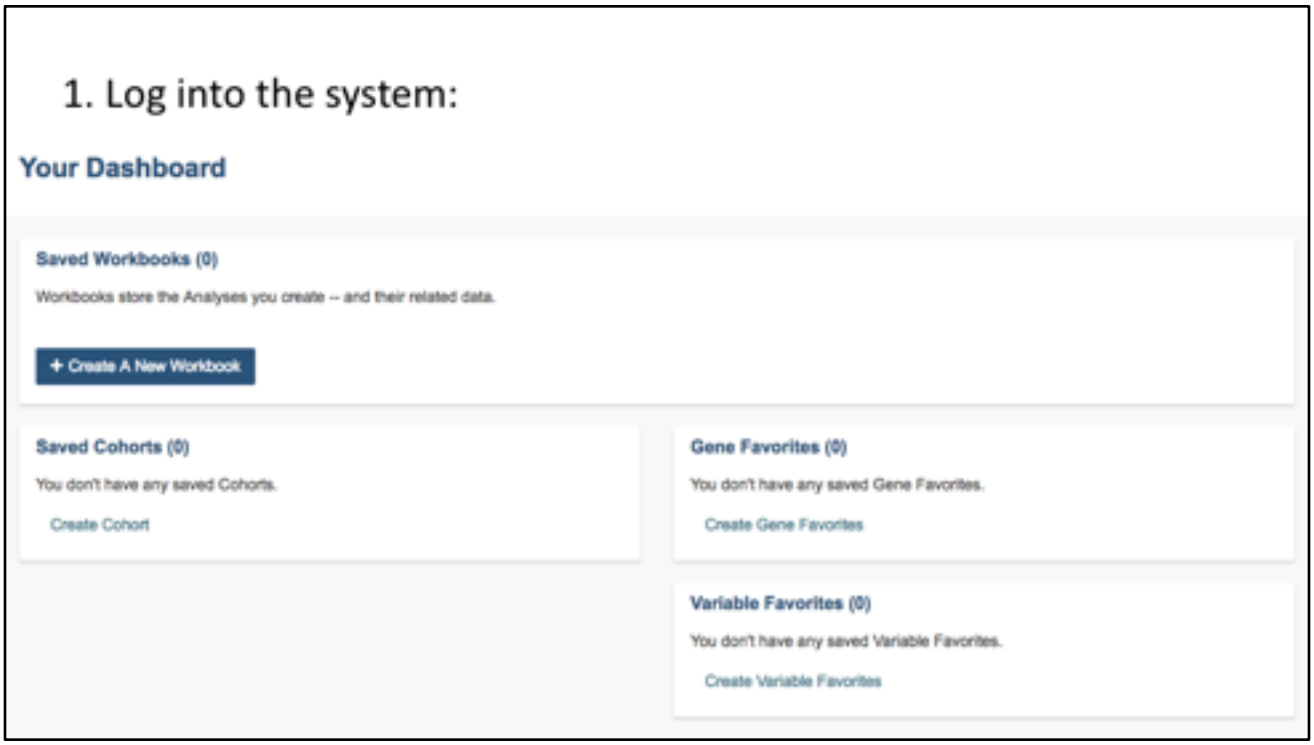
## The ISB Cancer Genomics Cloud

Institute for Systems Biology

Google Cloud Platform

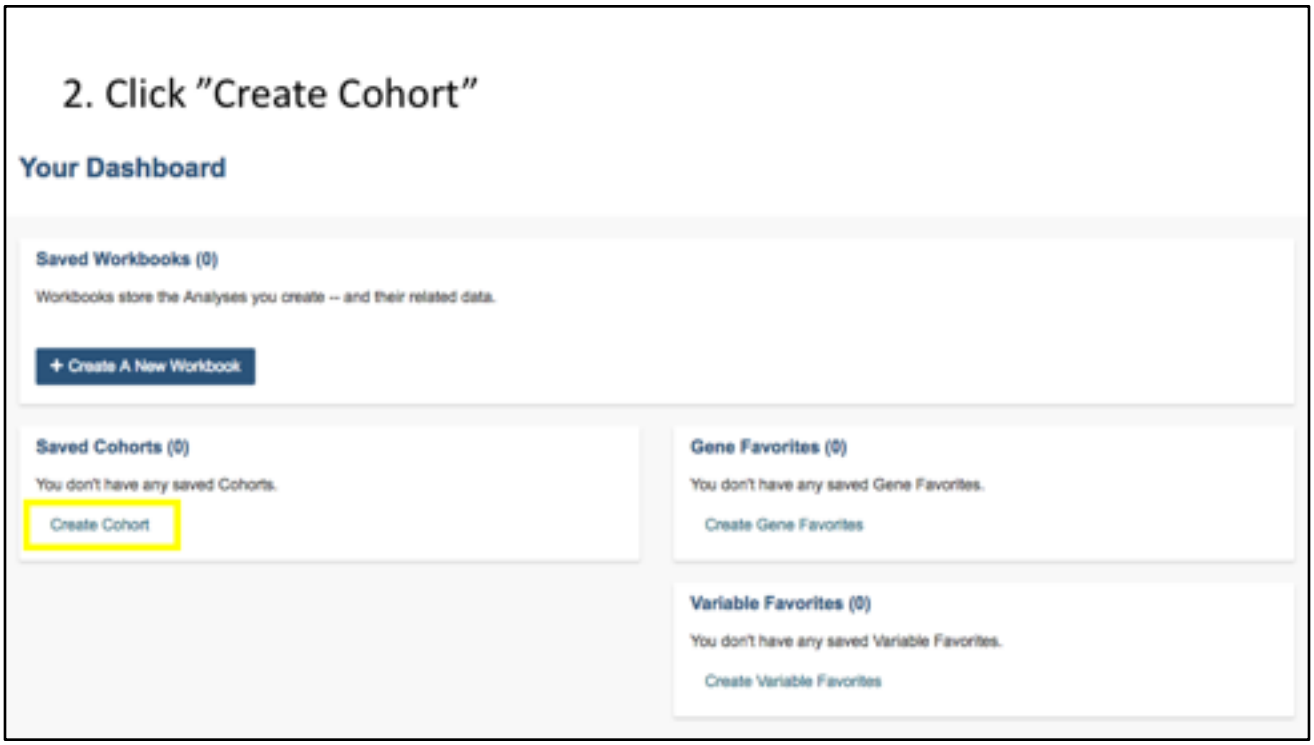
CSRA


- 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.

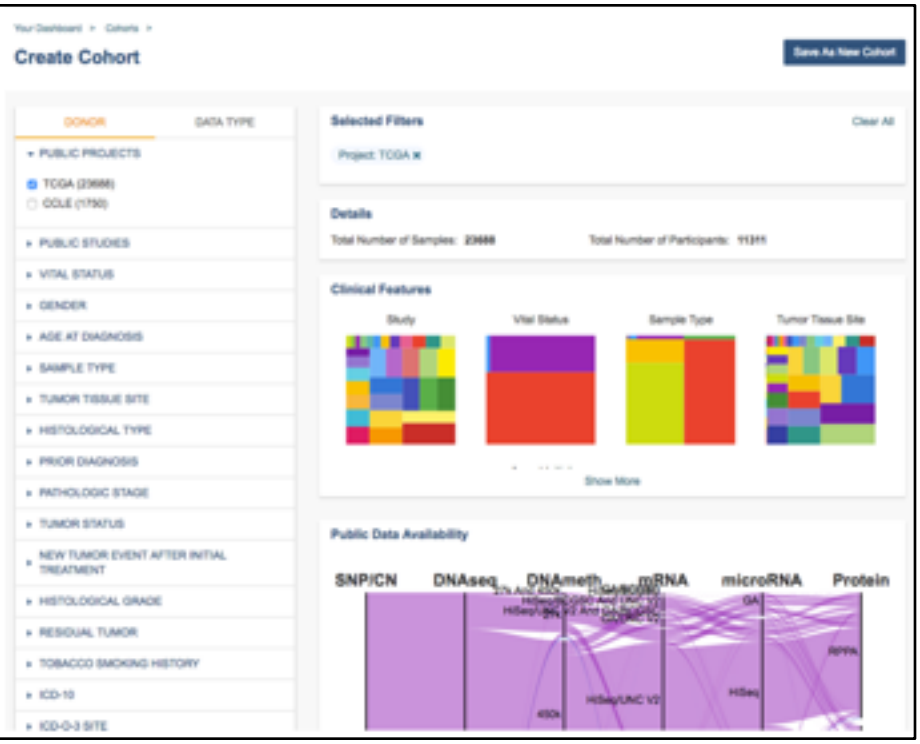



- After logging in, you are taken to the dashboard.
- This is where you can view an overview of the different workbooks and cohorts you create.
- Workbooks contain worksheets, where you can create analyses.
- Gene and Variable favorites is where you can define lists of interest to yourself.
- On top 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"

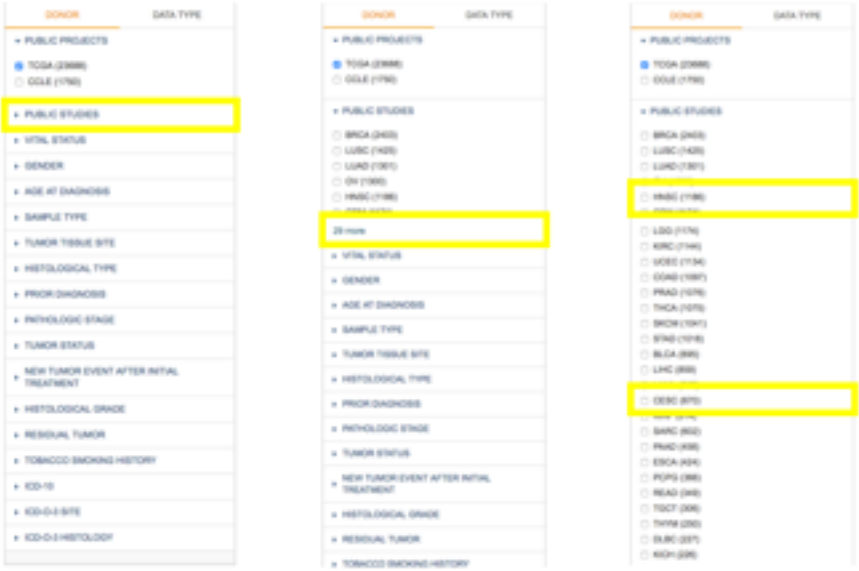


# 3. Cohort Creation




- On the left side, there are panels of features that you can use to define your cohort.
- The Details panel will show you how many samples and participants you currently have selected in your cohort. So initially we start with all of TCGA.
- The Clinical Features panel displays a visual breakdown of a few features of the current cohort you've specified.
- The Public Data Availability panel shows what kind of data is available for your current cohort.
- Note that there are two public data projects: TCGA and CCLE.

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



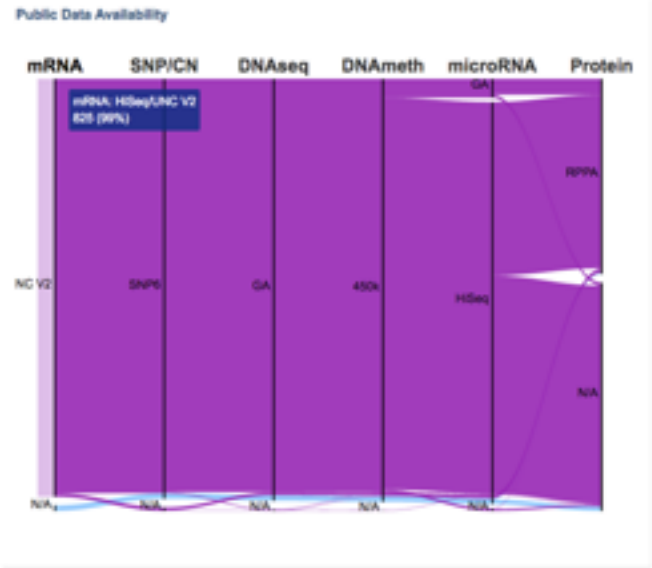
- For the purposes of our analysis, we will create a cohort comprised of all TCGA Head and Neck and Cervical samples.
- To do this we select those from the Public Studies.
- It is important to note that if we had not selected the TCGA Project, our cohort could include samples that are also from the CCLE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

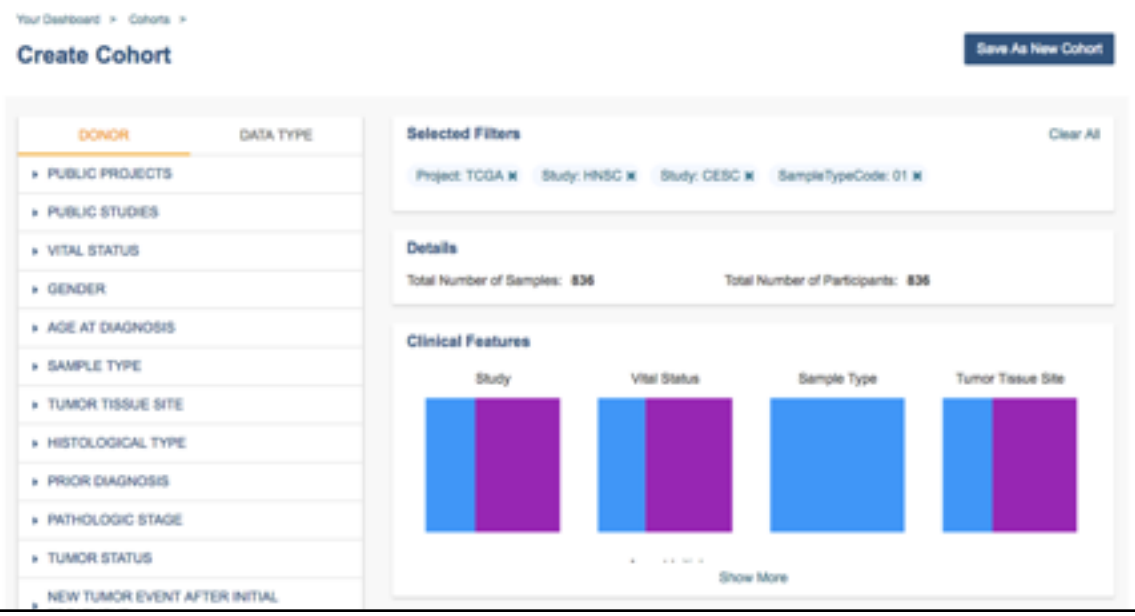
## 6. Select the Sample Type 'Primary tumor Tissue'




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

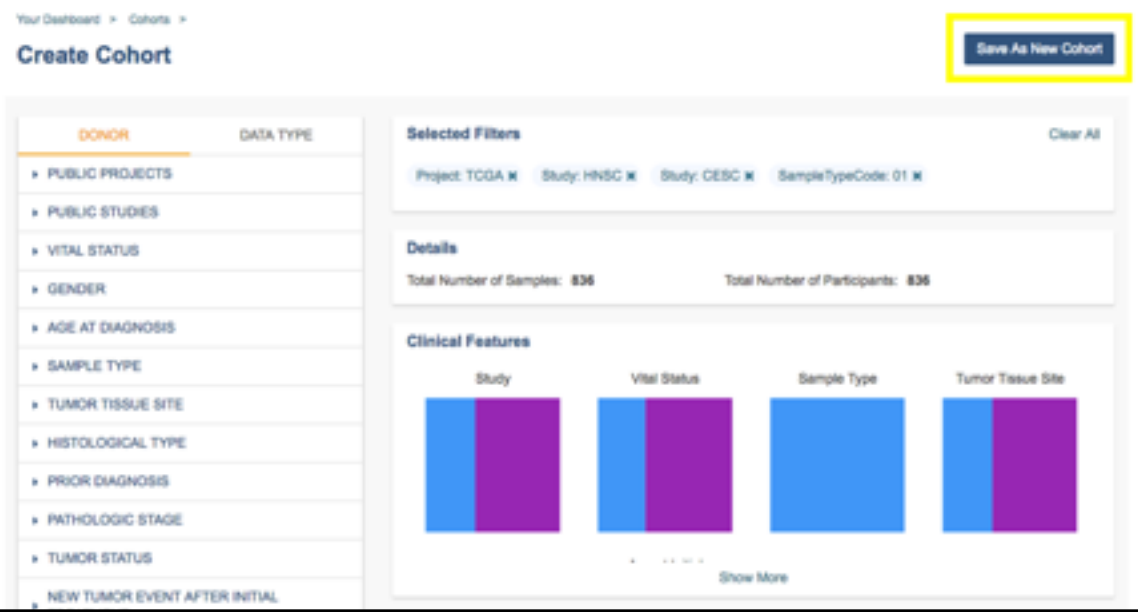
- After selecting only Primary tumor Tissue, we can see that most of our samples have gene expression data.

## 7. The resulting cohort




- 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




- 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

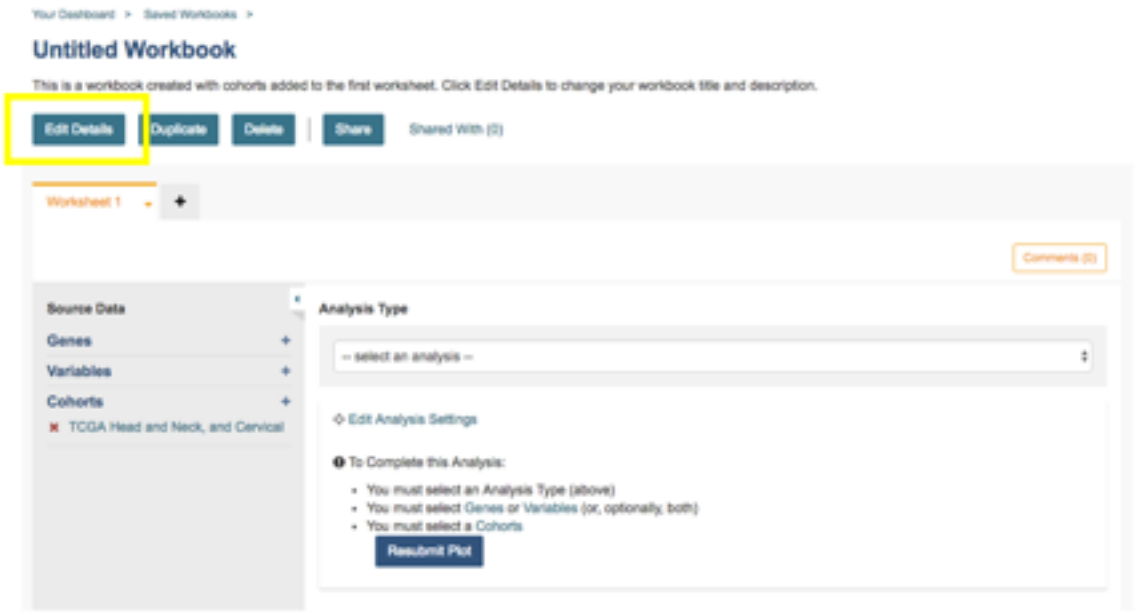
SAVED COHORTS | PUBLIC COHORTS

New Workbook | Delete | Set Operations | Share

| Cohort Name | # Samples | # Patients | Ownership | Shared With | Last Modified |
|---|---|---|---|---|---|
| TCGA Head and Neck, and Cervical | 836 | 836 | You | ( 0 ) | 05/18/2016 9:32 a.m. |
| TCGA HNSC CESC | 1744 | 836 | You | ( 0 ) | 05/16/2016 12:13 p.m. |
| TCGA HNSC | 1123 | 528 | You | ( 0 ) | 05/16/2016 12:12 p.m. |
| TCGA CESC | 621 | 308 | You | ( 0 ) | 05/12/2016 12:02 p.m. |
| DNA Meth | 213 | 214 | You | ( 0 ) | 04/12/2016 2:30 p.m. |
| GBM | 1174 | 605 | You | ( 0 ) | 03/14/2016 11:10 a.m. |

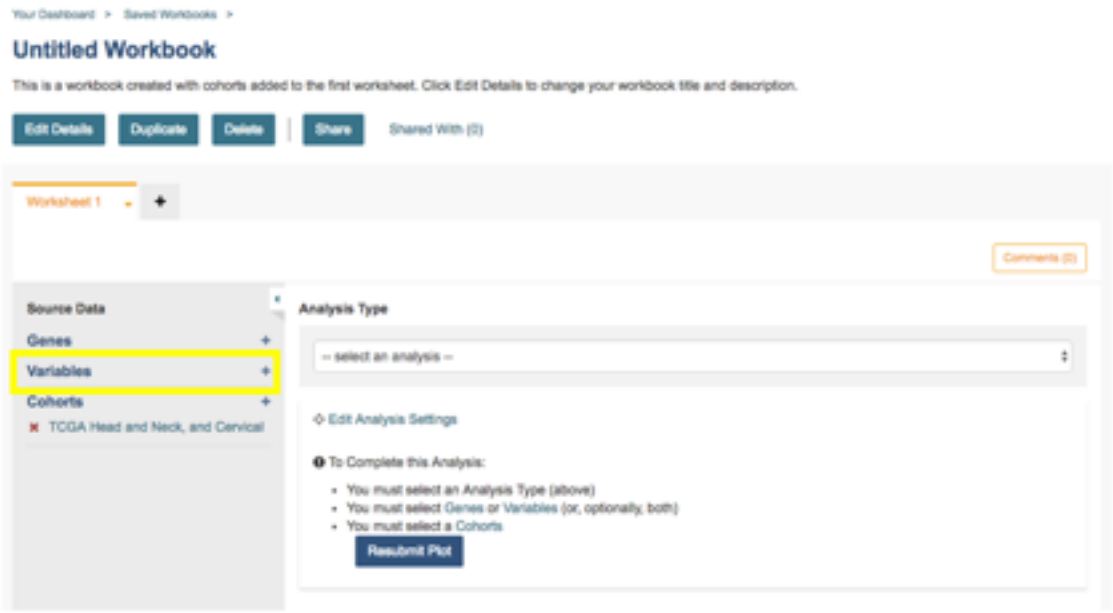
- This is where you can see all of the cohorts you've created and that have been shared with you.
- You'll notice that you also have access to Public Cohorts. These are cohorts that we've created for you. So far it's just one, but we plan on adding more.
- Another way that we could have created our cohort 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

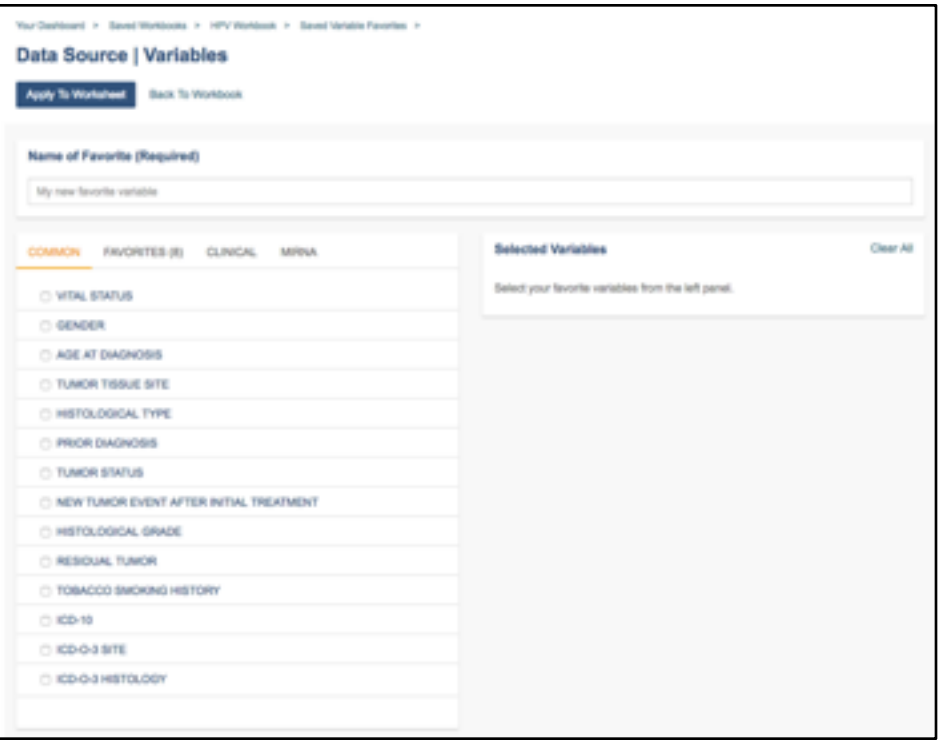



- When you create a new workbook, it is automatically populated with one worksheet.
- A worksheet is comprised of different data sources that you will use in your analysis. You can see that the Cohort we selected is already available.
- Let's first edit some details of our workbook by giving it a more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

## 11. Add Variables to your worksheet

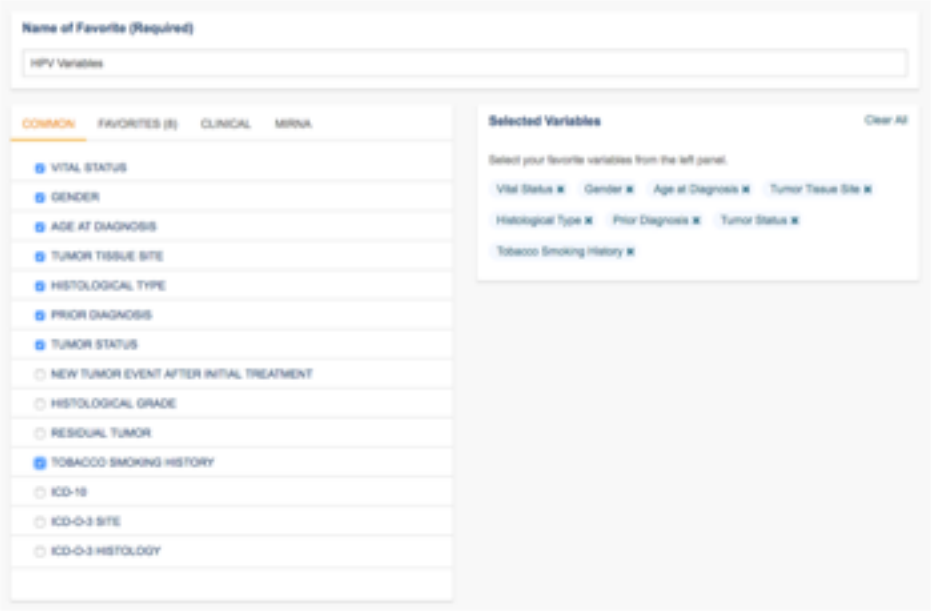


## 12. Creating a new Variable List



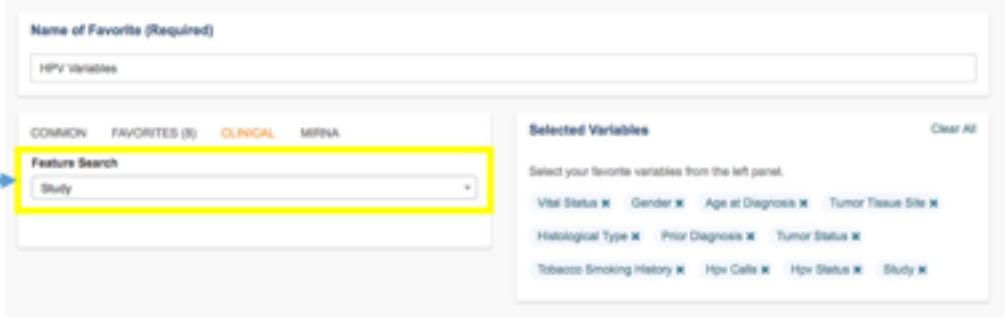

- If you don't already have variable lists created, you will be taken here. If you do, then you will be taken to the your list of previously created variable lists. To get to this page, click the Apply New Variable List button.
- The idea behind this concept is for you to be able to create a list of 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.




- 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




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 (0)

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 --

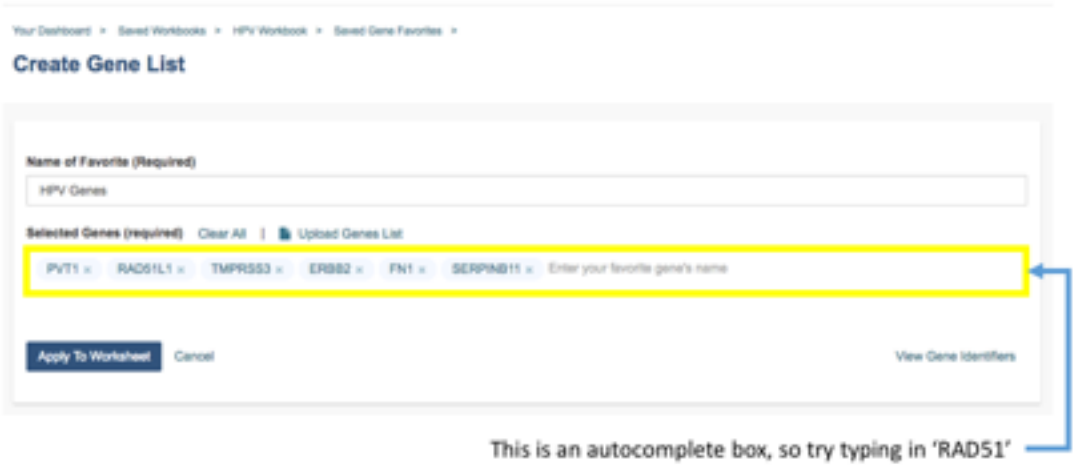
Edit Analysis Settings

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

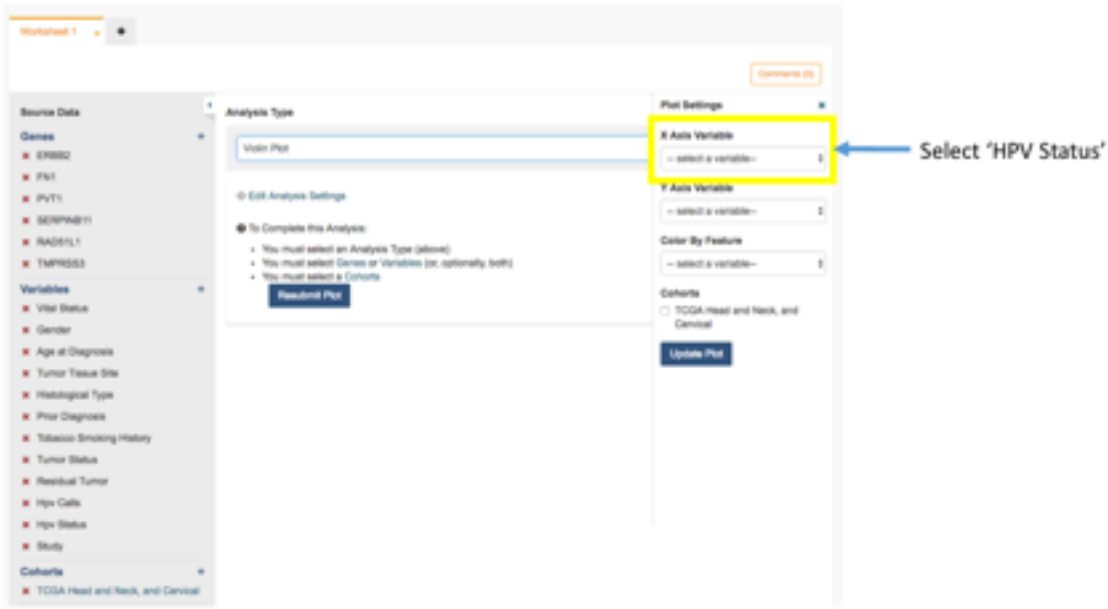
17. Create a gene list for your HPV analysis




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

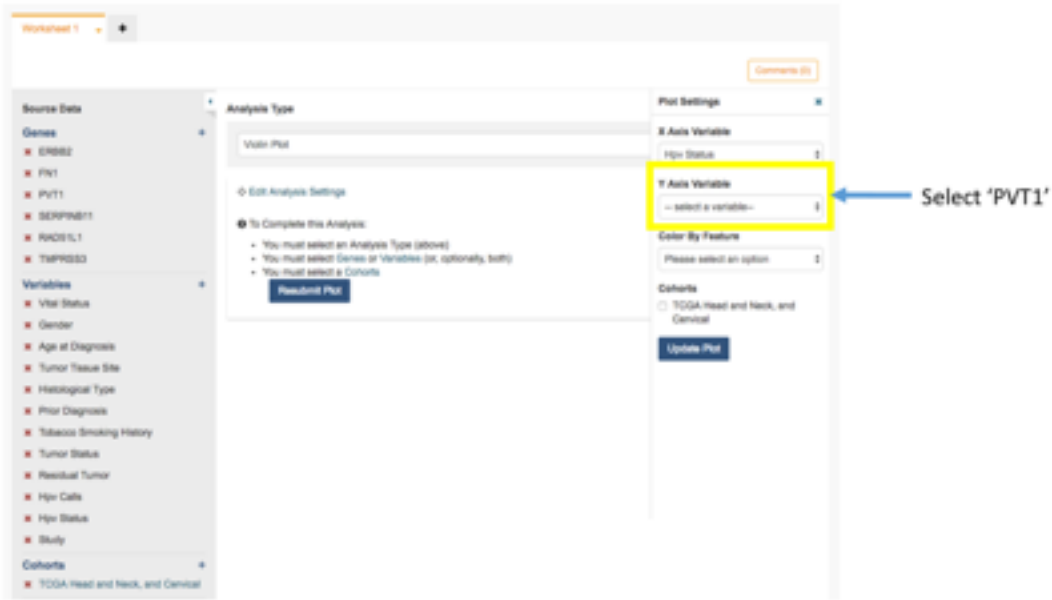
- Similarly to variables, if you have gene lists created, you will be taken to a listing of your gene lists.
- This page uses an autocomplete to help you find the genes you're looking for. Try typing in some of your favorite genes to see if they are in our system. If you're unsure of what your gene might 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



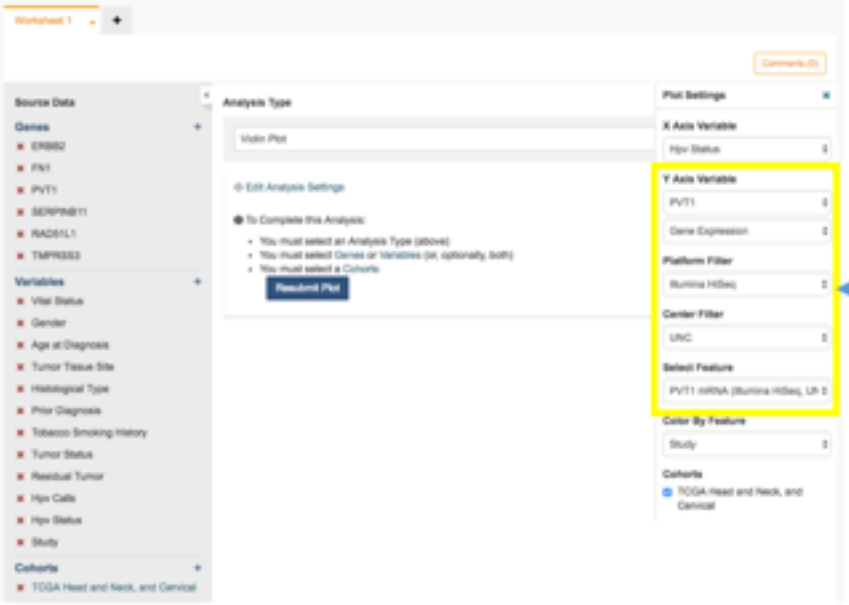

- Now that we have all our data sources ready, we want to start our analysis.
- We provide several different types of analyses (For more information please see our online documentation):
    - Barchart – 1 Categorical variable
    - Histogram – 1 Numerical variable
    - Scatterplot – 2 Numerical variables
    - 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




- Then we select PVT1 for our Y-Axis

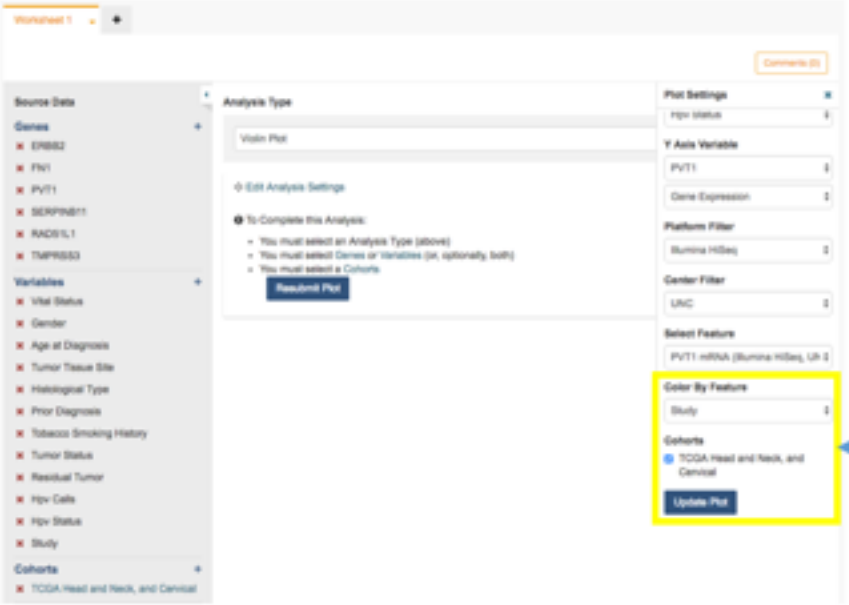
## 19. Creating a Violin Plot comparing HPV Status VS RAD51L1 Gene Expression




- 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.

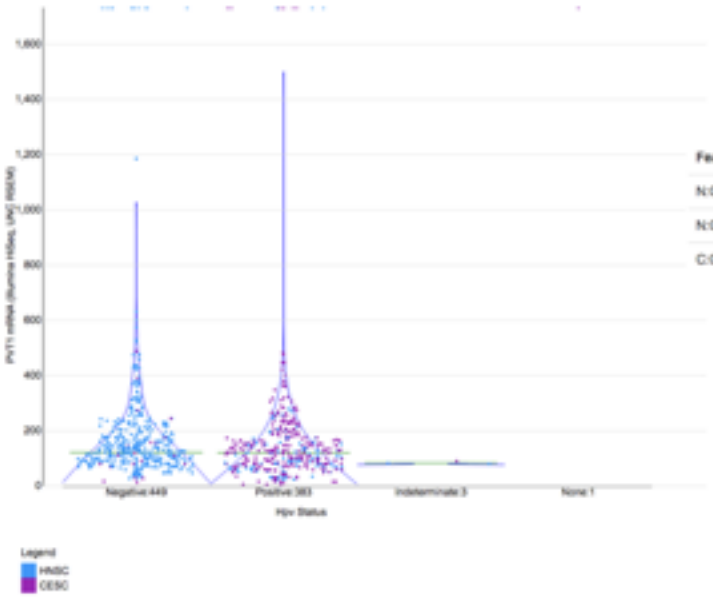
## 19. Creating a Violin Plot comparing HPV Status VS RAD51L1 Gene Expression




- 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:rnma_unc_illumina_hiseq | C:CLIN:hpv_status | 0.752 | 815 |
| N:GEXP:PVT1:rnma_unc_illumina_hiseq | C:CLIN:Study | 0.378 | 815 |
| C:CLIN:hpv_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.