

The Cancer Genome Atlas

Manual

John Loreti



Introduction:

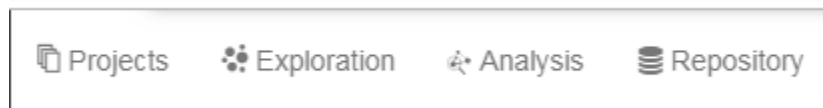
Cancer research is an ever-growing and changing industry. Each day new discoveries are made or studies are conducted, getting us one step closer in finding new treatment options and possible cures of this complex disease. In 2005, researchers wanted to conduct more research to gain a better understanding of genomic alterations that are associated with various cancer types. The information that was gathered was added to a database which is now known as The Cancer Genome Atlas (TCGA). By sequencing and analysing data from different cancers, researchers have been able to molecularly map over 33 different cancer types. Today, there are over 2.5 petabytes of genomic and epigenetic information on The Cancer Genome Atlas website.

The Cancer Genome Atlas is an important tool because it can be used as an aid in further research of different cancers and the information present on the database can be used in the development of medications or possible cures. Researchers can go onto the website to learn about a specific cancer, or they can submit their own research projects and findings. With the abundance of information that TCGA has to offer, researchers or bioinformaticians can even go into different projects and try to add on to the information that was collected from those studies.

Information found on TCGA:

- Gene names, symbols, location, brief description
- Research projects that are associated with specific genes
 - Within projects is information on:
 - Disease type
 - Primary site
 - Data (sequencing) and Experiments (tissue slides)
- Most frequent mutations
- Reference genome to perform additional/similar experiments
- Tools to compare data across projects

Major Tools



Projects: List of all of the various research projects on the database

Exploration: Observe specific cases within projects; separated by primary site, gender, project

Analysis: Compare survival rates, mutations, etc. across multiple projects

Repository: All of the files found on the database; from individual sequencing reads to simple nucleotide variations

The Cancer Genome Atlas Walkthrough

The best way to go explore all of the different features of The Cancer Genome Atlas is to go through using the database as a researcher/bioinformatician who had just discovered a gene that has not been linked to a type of specific cancer to gain more insight into the gene itself. Let's say that the scenario is that you are working in a laboratory and are asked to sequence and analyze a sample of tissue with a confirmed case of prostate cancer. In your analysis you come across some genes that are commonly associated with prostate cancer, but there are also some that have no previously known connections to prostate cancer. To find more information about one of

those genes, you search for it in The Cancer Genome Atlas. The sample study that is used is an article titled “The Genomic Landscape of Prostate Cancer”.

Links

The Cancer Genome Atlas: <https://portal.gdc.cancer.gov/>

Sample Study: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3709705/>

Spans L, Clinckemalie L, Helsen C, Vanderschueren D, Boonen S, Lerut E, Joniau S, Claessens F. The genomic landscape of prostate cancer. Int J Mol Sci. 2013 May 24;14(6):10822-51.

Part 1: Analyzing Data Using CGA

Step 1: Selecting and Searching for a Gene

After running the prostate cancer tissue sample, you decide to pick the gene NIPA2 from the experimental data to learn more information about it. On the main page of the Cancer Genome Atlas, go to the search bar on the left side of the screen and enter NIPA2 and search the gene.

The screenshot shows the NIH National Cancer Institute GDC Data Portal. The search bar on the left contains "NIPA2". Below the search bar, a list of results is displayed, including the gene symbol "NIPA2" and several sample IDs (e.g., 16e3d6cf-4eca-5f28-8928-d58c1bc6a631). To the right of the search results, there is a bar chart titled "Cases by Major Primary Site" showing the number of cases for various cancer types. The chart shows that Prostate is the most common site with over 4,000 cases, followed by Lung and Breast. The portal also displays the total number of samples: 440,782, 22,872, and 3,142,246.

Major Primary Site	Cases
Prostate	4,000
Lung	3,000
Breast	2,000
Colon Rectal	1,500
Stomach	1,000
Bladder	500
Esophagus	500
Head and Neck	500
Kidney	500
Soft Tissue	500
Testis	500
Thyroid	500
Uterus	500

Step 2: Analyzing the Search Screen Results

Under the “summary” section on the top left corner you are able to see the gene name, symbol,

type, location and description. “External References” sends you to other websites such as Ensembl to observe more data on the gene. Half-way down the page there are two charts which show the different projects where NIPA2 is included, and at the very bottom are the different somatic mutations associated with NIPA2.

Step 3: Exploring the Data

Right above the section titled “Cancer Distribution” is a button that says “Open in Exploration” which sends you to the “Exploration” section of the database which is an easier way to view all of the information the CGA has on the NIPA2 gene.

Step 4: Exploring the Data: Selecting a Project

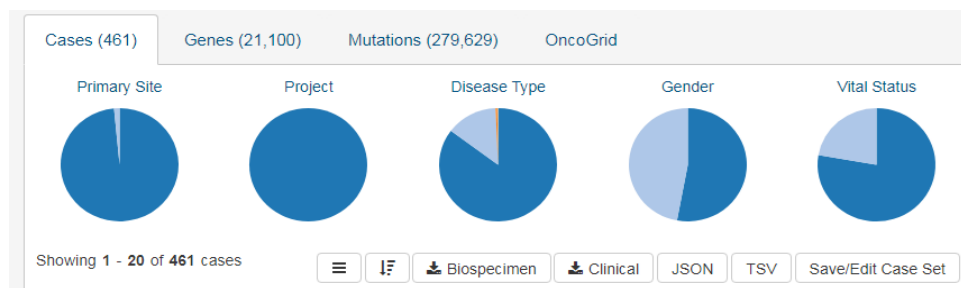
Let’s explore one of these projects to see one research experiment where NIPA2 was associated. Projects are listed by the different primary sites, or cancer sites, of the study. The first project that is listed is a project titled “TCGA-COAD”. Click on that project.

<input type="checkbox"/> Case ID	Project	Primary Site	Gender	Files	Available Files per Data Category							# Mutations	# Genes
					Seq	Exp	SNV	CNV	Meth	Clinical	Bio		
<input type="checkbox"/> TCGA-AA-3811	TCGA-COAD	Colon	Female	65	5	5	24	5	1	8	17	1	1





After clicking on the project, CGA brings you to a similar summary page when you searched for the NIPA2 gene, which is essentially a summary of the project. On the top right hand corner, click on the button “Explore Project Data”.

Step 5: Exploring the Data: Selecting a Project: Analyzing Project Data

From the next page you get something like this:



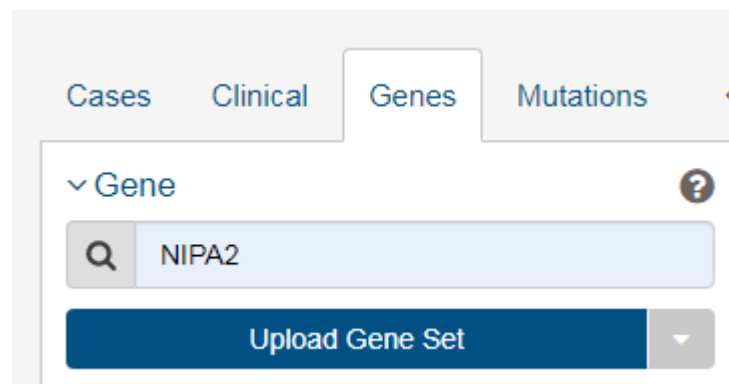
Here is where you can find the different information that was collected from this specific project. You can find the main primary site of this project, which in this case is the colon. You can also find the most common genes and most frequent mutations. Click on the “Mutations” tab to find the most common mutation in this project.

<input type="checkbox"/> DNA Change	Type	Consequences	# Affected Cases in Cohort	# Affected Cases Across the GDC
<input type="checkbox"/> chr7:g.140753336A>T	Substitution	Missense BRAF V600E	49 / 400  12.25%	565 / 10,202 
<input type="checkbox"/> chr12:g.25245350C>T	Substitution	Missense KRAS G12D	49 / 400  12.25%	208 / 10,202 



From the mutation tab we see that missense mutations on the BRAF and KRAS genes are the most common in this project.

Step 6: Exploring the Data: Finding NIPA2 Information Within a Project

To find how NIPA2 is included in this project, go to the top left search bar while you are in the “explore project data” page, click on the “genes” tab and search for NIPA2:



You can then once again go to the “mutations” tab but this time just see the most frequent mutations where NIPA2 is involved.

<input type="checkbox"/> chr15:g.22851821_22851822insA	Insertion	Frameshift NIPA2 L35Pfs*21	6 / 39  15.38%	16 / 10,202 
--	-----------	---	---	---

This information from various studies with NIPA2 mutations can be compared to the hypothetical study that you performed to see if any of the mutations are similar.

Step 7: Exploring the Data: Patient Profile

The patient profile from a study can also be compared to the different samples within a project.

Within the TCGA-COAD project, click on the Case ID TCGA-CK-6746. On this screen, you are able to view the different files, clinical data of the patient, biospecimen information, and also the most common mutations in this patient.

Step 8: NIPA2 Information from the Entire Database

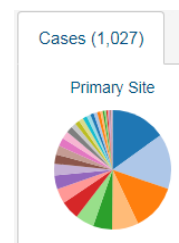
If you wanted to quickly get all of the information from NIPA2 that the CGA has to offer, follow steps 1-3 again but this time do not click on a specific project. From here click on the “mutations” tab to find the most frequent mutations of NIPA2 across the entire database, and you find that these two mutations are the most frequent:

<input type="checkbox"/> DNA Change	Type	Consequences	# Affected Cases in Cohort	# Affected Cases Across the GDC
<input type="checkbox"/> chr15:g.22851821_22851822insA	Insertion	Frameshift NIPA2 L35Pfs*21	16 / 944 1.69%	16 / 10,202
<input type="checkbox"/> chr15:g.22851830delG	Deletion	Frameshift NIPA2 G34Afs*11	15 / 944 1.59%	15 / 10,202

You can look through the entire list of mutations to try and match the mutation with the one from the hypothetical study to then research the data from the project where that same mutation was found.

Step 9: Exploring NIPA2 Data from other Prostate Studies

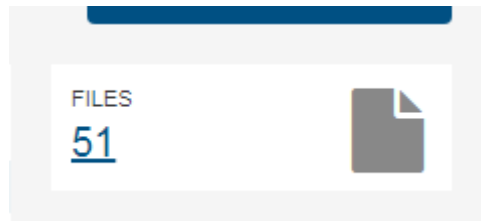
For your final analysis, let’s say you want to compare your NIPA2 to the information on the CGA specifically involving NIPA2 from the prostate. Go through steps 1-3 again, but this time on the pie chart under “primary sites” find where it says “prostate gland” and click on it.



From here you are able to find specific projects involved with NIPA2 and the prostate, where you can research and fully analyze the data that you collected from your study and compare that to what other researchers have found.

Step 10: Exploring NIPA2 Data from Other Prostate Studies: Patient Files

To get the files from a case within a prostate project, simply click on a specific case, similar to step 7. CGA brings you to a case summary page. On the top right hand corner is a place that says “Files”. Click on it.



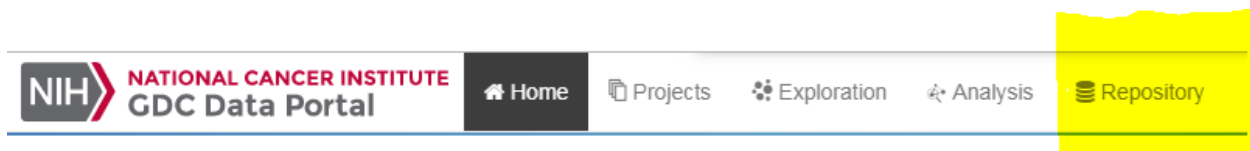
This brings to the repository where all of the files from this case are located. Some of the files, such as the sample nucleotide variation files are closed without special permission. However, some of the files are open to the public, such as Biospeciman and Transcriptome Profiling.

Part 2: Adding Research Data to CGA

The great thing about the CGA is that researchers are able to add the information that they have obtained from experiments to the database to give more information about different types of cancers. Here we will walk through the steps on how to get the reference genome that is used by the CGA to compare to the sequencing data from individual experiments.

Step 1: Finding Repository

On the main page from Step 1, click on the “Repository” tab on the top right corner.



Step 2: Analyzing a File

The repository brings you to a screen where every single file on the database is found. Some of the files say they are “controlled”, meaning that this file can only be accessed by having special permission. For the purpose of this walk through, that does not matter. When you get to the

repository, click on the first file.

Step 3: Analyzing a File: Finding the Reference Genome

When you click on a file, it brings you to a file summary page. Towards the bottom of this page, look for the “Reference Genome” section.

Reference Genome	
Genome Build	GRCh38.p0
Genome Name	GRCh38.d1.vd1

Step 4: Getting Reference Genome

Now that you know the name of the reference genome, GRCh38.d1.vd1, simply do a websearch of the name of the reference genome. The first web result is from the GDC website. Click on that link, and it then brings you to a downloadable link for the reference genome.

GRCh38.d1.vd1 Reference Sequence

GRCh38.d1.vd1.fa.tar.gz

- md5: 3ffbcbfe2d05d43206f57f81ebb251dc9

You now have the same reference genome that is used for the other projects on the CGA to use in your experiment to be able to add your own insight and experimental findings to contribute to the hundreds of files on the database.

Video Link

<https://www.youtube.com/watch?v=ICPr2inPR1Y&feature=youtu.be>