UCSC Genome Browser: coronavirus resources

The UCSC Genome Browser has been widely used by researchers working on human and other mammalian genomes for more than 20 years. The recent worldwide outbreak of coronavirus has inspired us to adapt the Browser for use by those wishing to understand the biology, epidemiology and immunology of the virus itself. Some display features have been built specifically for the coronavirus effort.

This tutorial is designed to familiarize virologists with the workings of the Browser and to show some of the data that are available. This is necessarily a brief introduction and shows only a small amount of the capabilities of the Browser, but it is designed to give you a way to think about the Browser and to get you started.

We'll start with the landing page, genome.ucsc.edu, and follow the coronavirus data link to the landing page for the coronavirus information.

[0:52] Covid-19 Portal

The information on this page is gathered from locations around the world and there is some basic information representing where the data come from and what kind of information is available.

To get to the Browser itself, click into this image on the left and it opens up a new window to the Browser. The Browser opens with the entire 29 kilobases of RNA showing in the window and all of the protein products.

[1:27] Browser paradigm

The basic paradigm is a set of coordinates representing the viral sequence and data "tracks" showing where sequences with biological significance can be found along that sequence. Anything that can be mapped to a location can be represented as a box on the Browser, which then can be clicked to access more information about that item. Genes, RNAs, proteins, binding sites, variants and other types of annotations can be shown on the Browser. It is a dynamic resource that allows the user to follow curiosity and test hypotheses against known information by quickly accessing the various data aligned to the appropriate places in the genome.

[2:09] Viral assembly names

You will notice that we named the SARS-CoV-2 virus assembly wuhCor1. This is not an attempt to blame the virus on the Chinese. It follows a long history of naming viruses after locations (such as Sendai and Ebola) and was chosen when we built the assembly very early in the outbreak. Once the engineering progressed, it was impractical and expensive to rebuild everything with a new name. We also use the official name of the virus throughout, which changed several times after we began building the browser.

[2:45] SARS-CoV-2 datasets

The Browser has a number of datasets available. Each dataset is represented in the Browser graphically as a "track" and a mouseover on the left side of the page shows the information about the track and kinds of data that are contained in it.

[3:02] UniProt tracks

Here you see the UniProt Processed Protein Products information and the next track down is identified as UniProt Highlighted Regions of Interest including the ACE2 receptor binding site on the spike protein of the virus. A number of other UniProt annotations are on by default: Signal Peptides, Transmembrane Domains, Disulfide Bonds. And other data tracks can be seen here as well.

Any one of these data tracks can be turned off using the right mouse button and choosing "hide" on the menu that comes on at that point. We'll turn off the Nextstrain track for now and come back to it later.

Oftentimes you can learn more about the individual items in a track such as the Protein Domains track, by clicking into the item and if you click into an item that is in dense visibility (the right mouse button here shows you the various visibility options), and switch it to "pack" you can see that these individual protein domains, as identified by UniProt, are made available.

[4:11] Details pages for annotations

If you mouse over an item in the Browser, you can see the longer name. If you click into it, you get into the details page, which gives you a lot more information, including, oftentimes, a reference back to the source article in the literature.

[4:29] Configuring track display parameters

Clicking into the little button on the left side of a track brings you to a page that outlines information about the track and some of the data conventions, including the coloring of the various tracks. And in this case, you get information about multiple UniProt tracks, including the Protein Domains track, the one we clicked into.

[4:53] NCBI Genes track

If I go back to the Browser now, I can go below the Browser graphic and view the number of different data sets that are available. The NCBI Genes track can be turned on from "hide" to "pack" by clicking into that and hitting "Refresh." And you can see here that these are full-length genes prior to processing, so it's a slightly different view of the viral genome than the UniProt Mature Processed Protein track gives you. Over here you can see the spike protein in the NCBI genes track.

[5:27] RT-PRC Primers

Another track of interest is found below the Browser graphic in the Mapping and Sequencing group. This is the RT-PCR Primers track. I'll turn it on to "pack" and hit "Refresh" and see here a number of primer sets that have been identified around the world in various locations and have been used to find the virus in biological samples. If you put the mouse over the track scale bar at the top of the Browser image and drag it to the right or left, you can zoom into a region, and you see here we've got a set of primers, including the CN-CDC_primer1 from China. You can see the direction of the alignment of the primer by the little chevrons, the little arrows, in the annotation on the Browser, and you can see that primer 1 and 3 point left to right, and primer2 points right to left.

You can also estimate the size of the viral genome detected by the primers by using the scale bar up above. You can see that it's roughly 120 bases between the two outermost primers.

You can also estimate it and mark it using the highlight feature, if you position your mouse in the top of the image and drag, the size of the selected region is shown next to the coordinate box, and you can leave a highlight in that place if you wish.

If you click into that item, then you can see the information about the primer. Its size is 21 nucleotides and it's a perfect match to the virus. If you click into the link, you can see the actual sequence of the primer itself and you can see the perfect alignment to the viral genome.

Let's zoom all the way out by a factor of 100 and that brings us back to the full size of the viral genome. We can close the primer track using the right mouse button, and hit "hide" and we'll also close the UniProt Domains track in the same fashion.

[7:29] Vertebrate-infecting Coronaviruses

Another track of interest is the Vertebrate Coronaviruses, and this is in the Comparative Genomics group and you can see that there are 119 vertebrate coronaviruses represented in the track. Let's turn the track on to "pack" and hit "refresh". You can see here that we have the alignment with all the variants in a large number of vertebrate coronaviruses that have been isolated from nature. You can see the virus that infects the bovine and the turkey. There are a few human coronaviruses in the alignment as well, including the virus from the 2003 outbreak.

It's possible to adjust the window on the side so you can see the full labels, by going into the Configure button at the bottom of the Browser and widen the label area. Let's set it to 30 characters so we can read a little better the actual names there. The names of the infected animals.

You can see that the alignment shows there is a lot of conservation but with variation and, as before, you can read the details by clicking into the button on the left side of the Browser and you can see all of the different sequences that have been incorporated: Anything with a checkmark is turned on in the current Browser display. Viruses that have been isolated from humans can be added to the display by checking the appropriate boxes in the upper section.

And below the text it describes the color conventions used in the track. In particular, it is good to notice that red blocks are drawn where there is a change in an amino acid and green where there is a variant that does not change the amino acid relative to the reference assembly. You can see that blue and yellow are described here as well. Yellow indicates that there is no sequence in the other virus with good alignment similarity.

If I go back to the Browser, then you can use that information to see that there are a number of regions where there is homology to the current SARS-CoV-2 coronavirus that infects humans, which is the base reference assembly against which all of these data tracks are aligned.

But there is a region over here, where there is a significant portion of the genome that has no homology with the viruses that infect human, except in these top two isolates, infecting bat and pangolin. If you go up to the Browser graphic here, you can see that the spike protein S1 aligns in this location.

So let's zoom into the region that encompasses the spike protein and you can see here that the bat and the pangolin viruses have sequences all through here where their sequence encoding the spike protein is a good match for the virus that infects human, although there are certainly regions where there is variation in the amino acid sequence as you can see with the red tickmarks on the aminoend of the protein. But these other viruses are significantly diverged and do not align well to the coronavirus that infects humans, the reference assembly.

We can zoom in even farther and look at some of the regions in the virus where the pangolin and bat matched the human-infecting virus, and you can see the amino acids along the top of the window at this point because we are now zoomed far enough in that we can see the actual sequence.

[10:51] PhastCons Conservation

The PhastCons Conservation track, when turned on to “full” gives you a sense for how well conserved the nucleotides along that region are with respect to the reference assembly.

If we zoom in even farther, then the nucleotide sequence is available. You can also see the amino acids in the non-reference isolates.

[11:15] Nextstrain Variants and Clades

Let’s turn off this track set and scroll down below the Browser graphic again and turn on the Nextstrain Variants and the Nextstrain Clades. These data are from a project that accumulates SARS-CoV-2 viruses infecting humans from all over the world and shows them in relationship to each other in a phylogenetic tree based on shared variants. These are the data that can be used to trace the transmission pattern of viruses from one place to another.

Let’s zoom out again using the ideogram up here and zoom out to the entire sequence of the virus, and you can see here the Nextstrain track which is updated multiple times a day, and shows you the various clades of the virus taken from samples from around the world. The data lines represent places where the isolate does not match the SARS-CoV-2 reference. The individual variants that define a clade can be seen graphically here where the isolates in a particular group have certain variants in common and certain subgroups have certain variants in common as well that identify them as being related to each other.

If you compare the colors on the left with the colors in the clades track above, you can see the relationship between the two. Let’s click into the B4 item here. And you can see here’s a list of samples belonging to this clade: from Shanghai, Australia, United States-Yale, Ghana, Russia, Netherlands, US-Wisconsin and, so these viruses are all related to each other according to sequence similarity and are displayed on Nextstrain.

Down below you can see information about those various clades, and the Methods section describes the methods used to collect and process the data as well. If you click into the nextstrain.org link itself, you go over to their website and you can see their representation of the data. Back to the Browser graphic.

If I click into the configuration button for this track, you see how I have many different subtracks available, only one of which is turned on at the moment. I can turn on the B4 track and configure the display and there are a number of display options. One option that's worth noticing is that if you increase the height of the display for this track, then you get enough room for the actual names of the variants themselves to be shown to the left of the graphic. So if I hit [submit], and then scroll down the page, you can see that this track is turned on and the relationships among the various isolates are seen graphically and their names are available to you as well.

Let's turn this dataset off by hiding it with the right mouse button.

[14:15] T-Reactive Epitopes

Below the Browser graphic in the Immunology section, let's turn on the T-Reactive Epitopes track. There are a number of different polypeptides here from two libraries M1 and M2. M1 is derived from the amino-terminus of the Spike protein while M2 is derived from the carboxy-terminus. These experiments identified regions of the spike protein using polypeptides from the 2003 SARS coronavirus that show cross-reactivity with T-cells isolated from COVID-19 in patients. Here we show those 2003 SARS peptide sequences aligned to the current virus, SARS-CoV-2. Let's turn them from "pack" to "dense" so can see just the epitopes themselves in the two groups.

[15:14] BLAT and Short Match

A useful feature of the Browser is the BLAT function. It allows you to find the location of any string of nucleotides of 22 bases or more (and to some extent shorter sequences as well). Let's get some sequence from one of the spike proteins to use as raw material.

We'll right-click into the small Spike protein S2' and Zoom to a window that has just that region, spanning about 1.4 kb. Then, in the top bluebar, choose "View, DNA," then [get DNA]. The virus is RNA, of course, but the Genome Browser was built for animals and the interface still says DNA. In several places, such as BLAT, and in the present case of obtaining sequence, the Browser uses T instead of U. We'll copy a bit of sequence, maybe 40 bases, go back one page, then select "Genome Browser" from the top bluebar and paste the sequence into the Position/Search box and [go]. A click on the "Browser" link returns us to the main page.   You see that we have navigated to a region defined by that sequence.

Let's go back and try it with a mismatch, which you might have, of course, if you are using sequence from an isolate you have sequenced yourself. I'll change a C to a T in two places and [go]. If I click the "details" link, I can see the side-by-side alignment showing the mismatches. If I go back one page and click "browser," I find myself zoomed to that sequence and can see the mismatches in the image.

There is also a Short Match track in the Mapping and Sequencing group that allows you to find all occurrences in your window of shorter oligonucleotides, down to 2 bases. Click on the link above the pulldown menu. If I use GGT, I find them all, on either strand. Note that the U nucleotide does not work in BLAT or in Short Match at the present time.

[17:33] Crowd-Sourced Data

A data track built specifically for the coronavirus effort is the Crowd-Sourced Data track. Because of the rapid discovery and dissemination of information about the virus, this track was built as a way for the latest information to be made public all in the same place as part of a very rapid release cycle. Anyone can release data in this track.

If you click on the link above the track control, you reach the configuration page, as with any track. However, for this track, you find a link that opens a Google doc that anyone can use to input information into the Browser.

As you can see here, there is a place to input the coordinates, labels, the source of the information, and your email address. The data are reviewed and released to the Browser during the night (US Pacific time).

Let's turn the track to "pack" and [submit].   There is an item in the current window. You see that the mouseover says, "evolution" and when you click the item, the details page shows a table reflecting the data from the spreadsheet, including a link to the original source of the information.

Here you can see the item in the spreadsheet that corresponds to the annotation we just saw in the Browser. Note the label "insertion" and the category "evolution," matching the label and the mouseover in the Browser graphic, respectively.

There are many other datasets available in the coronavirus Browser which we do not have the time to describe in this brief tutorial. And more data are being added all the time. In general, a good way to explore the data available in the Browser is to start with a mouseover on the track controls to see the long label, then click the link to read the documentation. If necessary, click through to the References and read the original papers. Turn on the track and have a look at the data.

So this has been a whirlwind view of some of the features of the Genome Browser and how to get around in the Browser. It is worth mentioning that the display features demonstrated here are universally available in the Genome Browser for any of more than 100 animals, without regard to which genome assembly you are using. The data, of course, will be different in each assembly.

We have more tutorials available on our video channel: https://bit.ly/ucscVideos, including a three-part introductory series, Genome Browser Basics.

Thanks for your interest in the Genome Browser and in the continuing effort to learn more about the coronavirus. Stay safe.