Using the UCSC Genome Browser Data Integrator

In this demonstration, we will show how to get variant data from the UCSC Genome Browser tables, using the Data Integrator.

We will begin at genome.ucsc.edu, the home page.

[0:15 Set up the Genome Browser]

We'll begin by setting the Browser to its defaults, starting with the Genome Browser link in the pulldown menu and '`reset all user settings`'. This sets the browser at the hg38 default genome assembly, and we will switch to the hg19 genome assembly for this demonstration.

The `go` button takes us to the browser graphic. Let's reset the browser by hiding all the data tracks, using the '`hide all`' button below the browser graphic, and then we'll turn on the UCSC Genes track to `pack.` Now we'll navigate to a genomic interval, which, when deleted on chromosome 22, represents DiGeorge syndrome. In the text box, we'll type in the gene coordinates, chr22:18,700,000-21,500,000, and hit the '`go`' button.

So we now have a display with 2.8 megabases of DNA on the page, and all the genes in that region.

[1:24 Set gene track to show single isoform]

Most genes have multiple isoforms, and we can reduce the display to a single isoform per gene by clicking on the little button on the left side of the gene track, going to the configuration page, and unchecking the box for splice variants. Now we're showing just a single variant for each gene.

[1:50 Display phenotype tracks]

Let's turn on a couple of data tracks now that represent information about the phenotype of the genes in question. We'll scroll down the page to the bluebar group, Phenotype and Literature, and we'll turn on the OMIM Alleles track to `pack`, and the OMIM Genes track to `pack`. We can also turn on the Gene

Reviews track to `pack`, and then the `refresh` button will give us these three tracks turned on.

We can see that there are only two genes for which there are authoritative reviews in the NCBI Gene Reviews compendium, but there are multiple genes that are annotated by OMIM, both at the gene level, the gene is relevant for phenotypic relationships and disease states, as well as a number of individual alleles for these various OMIM genes.

[2:40 Set up Data Integrator]

Now let's go to the tools menu and navigate to the Data Integrator using the pull-down menu. We'll see that the browser remembers what gene interval we were looking at. You also have the option to load in multiple regions, if you have multiple regions that you're interested in all at once that are not contiguous in the genome.

[3:03 Add datasets to Data Integrator]

We'll begin by adding data sources, and we'll start with the track group, Genes and Gene Predictions, and we'll add the UCSC Genes track to the list of tables that we're going to interrogate using the Data Integrator.

Now let's add another data set from Phenotype and Literature group. These track groups reflect the same structure as the bluebar groups on the main page, back at the Genome Browser graphic. We'll go to track, OMIM Genes, add that, and track, OMIM Alleles, and add that dataset.

[3:41 Choose data fields to export]

We will use the `choose fields` button to allow us to pick out which bits of data from these three tables we're interested in, as well as some related tables.

We'll start with UCSC Genes and we'll simply clear all of these data fields.

We'll add in the related table, the KG cross-reference table [kgXref], and we'll select chrom, transcription start [txStart], transcription end [txEnd] from the knownGene table, and geneSymbol from the KG cross-reference table, and for OMIM Genes, let's just get the name of the gene and not repeat the chromosome coordinates. We'll relate it to the table omimPhenotype, add that table to our list, and we will choose a description which gives us a phenotypic description in a small number of words for the individual genes in our query. And then for the OMIM Alleles table, let's just get the chromStart as these are typically single-nucleotide alleles and chromEnd will be obvious; and we'll get the name. Finally the `Done` button at the bottom takes us back to the main page and we are ready to `get` our `output`.

We simply `get output` and dump the data to the screen, which is suitable for porting it to programs or some kind of a script if you want to mine the data.

[5:12 Load data into spreadsheet]

Let's go back to the Data Integrator and load it directly into a spreadsheet by sending output into a file.

We can call our data "variantData," and at `get output` we're now asked to save the file. It will load each column into columns on the table using tabs as the delimiter; and so each of the columns from the various tables will be incorporated into the spreadsheet into a separate column.

I'm going to double-click here at the top of Column F to expand the omimGene2.omimPhenotype_description field so that you can see all of the descriptions that are reflected in the data.

That concludes our demonstration of how to export data using the Data Integrator. This tool can be used for integrating data from any one of the data tables in the genome database. It's possible to add data from up to five different primary tables in the Browser. Plus, as we have already demonstrated, many of the secondary tables as well.

Thanks for watching and thanks for using the UCSC Genome Browser.