dbSNP resources in the UCSC database

Today we will discuss some of the variation data from dbSNP as displayed on the UCSC Genome Browser.

We will start at genome.ucsc.edu at the main Browser page and we will 'reset all user settings', which brings us to the most recent genome assembly, hg38.

[0:28 Set up GB to hg18 defaults]

So to begin, let's go down to hg18, one of the earlier version of the genome, and we will hit the *go* button. At the Browser graphic page, we will hit the 'hide all' button to turn all the data tracks off.

[0:45 Turn on SNPs (130) track]

Scrolling to the bottom of the screen, to the Variation and Repeats bluebar group, we'll draw your attention to the dbSNP track, SNPs version 130. In addition, we have earlier versions: 129, 128 and 126. At the time of version 130, before moving to hg19, dbSNP was mapping all new SNPs to the hg18 genome assembly.

If we click into the link above the pulldown menu, we can see that this table was last updated in 2009. We're going to turn the track on to 'dense' and then 'Submit'. This shows us all the SNPs in version 130, and let's turn on the gene track, UCSC Genes to 'pack' so we will see the scale. We have 310,000 bases and we have the default gene GABRA3. The region is annotated by many SNPs. Now let's zoom into one of the exons here in the middle of the screen, and you can that at this resolution, around 5,000 basepairs, there are a dozen or so SNPs in view.

[2:02 "Polymorphism" and why hg19 has SNP subsets]

Around the time of snp130 we learned something about the SNP track and the way people were using it that gave us some concern. Researchers were filtering their datasets with the contents of the SNP data, under the impression that they were common in the population and that such filtering would leave behind rare

variants in their samples. You will notice is that we use the term "polymorphism" to describe the track, which to a geneticist implies a variant that is fairly common. Let's open a new web browser window and go to Wikipedia, and type in 'Polymorphism'. This is certainly not the most scientific of sources, but you get a fairly decent look at the consensus view of what polymorphism means, and you see here on the disambiguation page that the Wikipedia definition carries with it the phrase "alleles, … each with an appreciable frequency in the population." Reading in other sources indicates that the threshold is around 1% minor allele frequency.

Let's go back to the Genome Browser. When we learned that people were using this definition of "polymorphism" to filter out what they believed were common variants from their datasets, the UCSC team decided to help make it more clear: The dbSNP database contains essentially =any= variant that has been detected and submitted, without regard to how common it may be in the population. If you take the SNP 130 track and filter your sample with it, you are filtering out a lot of rare variants as well, so the word "polymorphism" is probably inaccurate in this context. Nevertheless it has become commonplace to refer to a single-base variant as a SNP. In SNP tracks on hg19 and thereafter, the common SNPs are separated by UCSC, as we shall see.

[3:54 Why "Simple" Nucleotide Polymorphism?]

You'll also notice something else about the nomenclature. The Genome Browser uses the word "simple" instead of "single" in the acronym for SNP, primarily because NCBI had been using it that way; some the variants in the database are larger than a single nucleotide. So let's navigate to the hg19 assembly, the next most recent version of the genome assembly. And we will 'hide all' the tracks again. We will scroll down the page again. On this assembly we have separated the Variation from the Repeat section.

[4:35 All SNPs track and subset Common SNPs on hg19]

At this time you will see that on hg19 we have a track called Common SNPs, version 147 as the first track in the Variation group. Let's turn it on to 'dense'. You will notice that there are other Common SNPs tracks, 146, 144, 142,141 and 138. The Common SNPs tracks are all tracks that have been filtered from the

corresponding All SNPs track. These Common SNP tracks are tracks that really *are* common in the population, using the accepted definition of 1%. In addition, we do still have the All SNPs tracks, which are the complete set of SNPs, analogous to the snp130 track we saw on hg18.

There are two other kinds of SNP tracks that represent different subsets of the All SNPs tracks, the Flagged SNP track and the Multiply Mapping SNPs. We'll set these to 'dense' as well. Briefly, the Flagged SNPs are those that have been flagged by dbSNP as having been contributed by a locus-specific database, for which it is possible that a phenotype might be associated. In the dbSNP database, the "clinical bit" is set to 1 for these SNPs. These SNPs are typically in a gene for which variants have been shown to be phenotypically significant. A SNP in the Flagged SNPs track is not necessarily a risk allele itself, however, but it is in a gene for which there are risk alleles.

The Multiply mapping SNPs are those that map, using the flanking sequences around the central variant, to more than one place in the genome, so it is good to know about those if you are doing a genome-wide association study.

[6:21 Compare SNP tracks]

So let's refresh and load these tracks into the Browser. We can see that the number of SNPs in these tracks varies widely among the different tracks. All SNPs has a very high density because it has all of the SNPs from the dbSNP database regardless of how common they are. It is a proper superset of the other three tracks. The Common SNPs are essentially SNPs that are found, as you can see from the mouseover, in 1% or more of samples. In addition, they have to be present in the database at least ten times, to avoid treating certain rare variants as common based on dbSNP's reported minor allele frequency alone. Some SNPs in dbSNP that are represented as 50% minor allele frequency, but they are cases where the variant has been reported from only one person ever, as a heterozygote. In that case, the minor allele frequency is dutifully reported by NCBI as 50.000%, but it is *not* a common variant!

You'll notice that in this region there are no multiply mapping SNPs and that Flagged SNPs are represented in little clusters of red, and you'll see that when we

turn on the UCSC Genes track, those SNPs will line up with the exons, in this case with the SOD1 gene, which is the default location for the hg19 assembly.

[7:45 SNP track configuration page and schema]

So let's scroll back down to the bottom of the screen and look at the Common SNPs again and look at a couple of things on the configuration page. If we click into the link above the pulldown menu, we see first of all that this is a 2016 July release, and we'll also find that if we click into the 'table schema' link, we can learn how many variants there are in the table: nearly 15 million.

Each of the SNPs tracks (and indeed all tracks on the Browser) has a description that gives details about the construction and interpretation of the track. You see here that the four tracks are described as is how they relate to each other. There is a large amount of other information on the Description page, including interpretation of the color scheme in the display, so you can see why the SNPs in the Flagged SNPs track were red, because each variant is likely to disrupt a protein: coding-nonsynonymous, such as missense or nonsense or they are splice-site variants.

You can see that green is coding, non-synonymous, and that black and blue are not in protein-coding regions. A lot of other information is available about these tracks farther down the page.

Navigating back to the top of the page, you see that there are filtering options if you wish to view only a certain type of SNP, and also coloring options if you wish to highlight a certain type of SNP in its own color or if you are colorblind and perhaps have difficulty distinguishing green from red.

[9:26 All SNPs table schema and row count]

We now go to the Genome Browser using the link in the top bluebar, and remembering that there were nearly 15 million SNPs in the Common SNPs track, let's look at the All SNPs track. This time, let's click into the little button to the left of the track, which actually takes us to the same configuration page as does the link in the track controls below, and you will notice that this table was loaded on the same day in July 2016, as was the Common SNPs track. And clicking into

the 'table schema', you will see here that this table has more than 154 million SNPs. So it has more than ten times as many as the Common SNPs table. In the full dbSNP dataset, common SNPs are much less common than rare SNPs!

[10:12 A single SNP and its details page]

Now let's go to a particular SNP and have a look at the details page. We want to use rs2507733 and then navigate there using the 'go' button and we will choose the most recent SNP tracks, 147, and you will see that our SNP is highlighted in the middle of the page. So let's click into that SNP and have a look. You can see that for this SNP the observed alleles are C and T. The genome reference has a C for human and a T for chimp, orangutan and macaque. These are the ancestral alleles. You can see that if we scroll down the page that it has been contributed by a large number of different sequencing projects, and in fact, if you look at the number, the SNP has never been seen since the original reference assembly was established: No other occurrences of the C and 126,000 occurrences of the T in homozygous samples. So the reference at this location is really a private SNP; it has not been seen since the sequencing of the reference, and the T that you see in chimp, orangutan and macaque is really the ancestral allele as well as the major allele in the human population.

[11:30 A change between genome assemblies]

So let's go back the Browser display using the Genome Browser link at the top, and we'll zoom into this region. And we'll remove the highlighting using the right mouse button. Let's use "View in Other Genomes" to jump to the homologous region in the hg38 genome assembly. If we submit here for hg38, we have a perfect match, 100%, and you see that we still have the same reference SNP highlighted in the middle of the range. Clicking into the SNP on hg38, you see that the alleles are still C and T, but the reference is now recorded as a T instead of the C we saw on h19. That means that on hg38, the Reference Consortium has replaced the original with a new clone that carries the common allele. You can see that the C is still 0% and the T is 100% and the allele counts have the same numbers as before. The C in the original genome assembly is not counted, presumably because no one has submitted it as a variant to dbSNP.

Returning the Genome Browser, we will turn on the Hg19 Diff track to pack. This is the track where we record on the Genome Browser which regions are different between the hg19 and the hg38 genome assemblies. You can see that the item in this track is colored red, and if we click into that item, we learn that a "new contig has been added to update sequence or fill gaps present in hg19." You can see in the item-specific details above that the new contig added to the hg38 assembly is 333 bases long.

Essentially what has happened is that the GRC has pulled a different clone out of the freezer, sequenced that, and substituted it back into the tiling path in the reference. Because the reference consortium wishes to have a solid relationship between any sequence in the reference and an actual accessioned clone, they are not going to simply substitute a new nucleotide without annotating the accession numbers of an actual molecule that you could order and sequence yourself.

[13:58 A Common SNP details page]

Clicking back to the Genome Bowser now, you'll notice that this region has all four of the SNP tracks turned on and that in this location only the All SNPs track has an annotation. It makes sense that in a coding region, the Common SNPs are not too abundant, because a SNP in a coding region is often deleterious and therefore eliminated from the population. You can see that there is one SNP in the region that represents a variant that is a true polymorphism, at 18% in the population, and 81% reference.

You will also notice that in only one of the isoforms is this variant annotated as an amino-acid change from a proline to a serine. It is not a coding region in the other isoforms. If we click into the 'Position' link at the top of the details page, then zoom out by 10x, we see that the proline is at the amino acid 3 on the amino end of the protein.

We can zoom out by a factor of 100 and you notice that All SNPs and Common SNPs tracks are more populated. Using the right mouse button, let's hide the GTEx expression track. You see in the Hg19 Diff track the full extent of the clone colored red and you see clones on either side of it that show some modification in hg38 relative to hg19 as well.

[15:24 Two ways to navigate between genomes]

One thing you may have noticed in navigating around the Genome Browser is that we have used different methods of moving from one genome to another. Up here in the "Genomes" pulldown, you can see that we can get to hg19 by clicking on that choice in the menu, and that also in the "View" menu, we could "Convert" and go to hg19 as well. The difference between these two methods of travelling between genomes is that in the latter case we are using homology to get you to essentially the same place in the other genome and in the Genomes pulldown, we are remembering the last place you visited in that genome and we are going there.

To demonstrate, let's navigate to some other gene in hg38. Let's just choose FGFR2, navigate over to FGFR2 and you'll see that we have all the same tracks turned on and we are in a 118 kilobase region. If we go now to "Genomes" and click on hg19, it should go to the last place in hg19 that we were, not to the FGFR2 gene. And we are at the PIGP gene. We can jump back to hg38 using the "Genomes" link and we are back at the FGFR2 gene on hg38.

However if we go to hg19 using the "View > In other genomes (Convert)" choice in the other pulldown menu, and go to hg19, we can expect to be at the FGFR2 gene because we are using genomic homology to find the orthologous position in the other genome. So there are actually two ways to get back to a previous assembly, to the place where you were before (using the Genomes menu) and the place that matches by homology to the place where you are at the moment (using the View menu).

Thanks for watching our tutorial today and thanks for using the UCSC Genome Browser.